DOCUMENT RESUME

ED 400 713	FL 024 200
AUTHOR	Brown, James Dean, Ed.; Yamashita, Sayoko Okada, Ed.
TITLE	Language Testing in Japan.
INSTITUTION	Japan Association for Language Teaching, Tokyo.
REPORT NO	ISBN-4-9900370-1-6
PUB DATE	95
NOTE	193p.
AVAILABLE FROM	JALT Central Office, Urban Edge Bldg., 5th Floor, 1-37-9 Taito, Taito-ku, Tokyo 110, Japan.
PUB TYPE	Reports - Descriptive (141) -- Guides - Classroom Use - Teaching Guides (For Teacher) (052) -- Collected Works - General (020)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	Behavioral Objectives; Classroom Techniques; Cloze Procedure; College Entrance Examinations; College Freshmen; Comparative Analysis; Criterion Referenced Tests; Decision Making; English (Second Language); Foreign Countries; Higher Education; Industry; Language Proficiency; *Language Tests; Nonverbal Communication; Norm Referenced Tests; Oral Language; Program Development; Pronunciation Instruction; Second Language Instruction; *Second Languages; Standardized Tests; *Test Construction; *Test Use; Test Validity; *Verbal Tests
IDENTIFIERS	ACTFL Oral Proficiency Interview; *Japan; *Oral Proficiency Testing; Teaching to the Test

ABSTRACT

Papers on second language testing in Japan include: "Differences Between Norm-Referenced and Criterion-Referenced Tests" (James Dean Brown); "Criterion-Referenced Test Construction and Evaluation" (Dale T. Griffe); "Behavioral Learning Objectives as an Evaluation Tool" (Judith A. Johnson); "Developing Norm- Referenced Tests for Program-Level Decision-Making" (Brown); "Monitoring Student Placement: A Test-Retest Comparison" (Sayoko Okada Yamashita); "Evaluating Young EFL Learners: Problems and Solutions" (R. Michael Bostwick); "Good and Bad Uses of TOEIC by Japanese Companies" (Marshall Childs); "A Comparison of TOEFL and TOEIC" (Susan Gilfert); "English Language Entrance Examinations at Japanese Universities: 1993 and 1994" (Brown, Yamashita); "Exploiting Washback from Standardized Tests" (Shaun Gates); "Testing Oral Ability: ILR and ACTFL Oral Proficiency Interviews" (Hiroto Nagata); "The SPEAK Test of Oral Proficiency: A Case Study of Incoming Freshmen" (Shawn Clankie); "Making Speaking Tests Valid: Practical Considerations in a Classroom Setting" (Yuji Nakamura); "Cooperative Assessment: Negotiating a Spoken-English Grading Scheme with Japanese University Students" (Jeannette Mclean); "Assessing the Unsaid: The Development of Tests of Nonverbal Ability" (Cloze Testing Options for the Classroom" (Cecilia B. Ikeguchi); and "The Validity of Written Pronunciation Questions: Focus on Phoneme Discrimination" (Shin'ichi Inoi). (MSE)

ED 400 713

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

James Dean Brown and
Sayako Okada-Yamashita, Editors

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Gene van Troyer
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

FL 024 200

2

The Japan Association for
Language Teaching

COMING SOON!

On JALT 95: Curriculum and Evaluation

Proceedings of the 21st JALT International Conference on Language Learning/Teaching

So you wanted to come to the 1995 JALT International Conference, but for some reason you just couldn't make it. What happened there? What did you miss that's professionally important to you? How can you find out?

In the past you couldn't get the answers to these questions, except through word-of-mouth or less than satisfying presentation reviews that would appear in JALT's monthly magazine, *The Language Teacher*.

This year that's going to change. The answers (at least some of them) will be in *On JALT 95: Curriculum and Evaluation*, the proceedings of the 21st JALT International Conference on Language Teaching/Learning. Finally, a book that addresses the events and concerns of the members of Asia's largest language teaching professional organization–the Japan Association for Language Teaching. *On JALT 95* will feature the plenary addresses and speeches of the Conference's invited guests, in addition to articles based on the broad spectrum of colloquia, roundtable discussions, and presentations made at the 1995 conference in Nagoya.

This volume is a "must read" for anyone interested in language teaching and teaching research in Japan and Asia in general. It is the most comprehensive collection of articles in English that has appeared on FLT in Japan in two decades, and it will be sent to all Conference attendees as part of their registration package. If you weren't an attendee, here's how to order a copy:

Price*: (*in Japan*) **¥2500**. Orders should be sent by postal transfer (*yûbin furikae*) using the form at the back of *TLT*, or a regular postal transfer form bearing the account **#YOKOHAMA-9-70903-JALT.** In the message area be sure to write "JALT 95 Conference Proceedings." (*International*) **¥3500**. Checks or international postal money orders must be made out in the proper yen (¥) amount. Sorry, no credit card orders accepted. Orders should be sent to:

JALT 95 Proceedings, JALT Central Office
2-32-10-301 Nishi-Nippori, Arakawa-ku, Tokyo 116

On JALT 95: Curriculum and Evaluation **ISBN: 4-9900370-1-6**
Published by the Japan Association for Language Teaching

***NB**: *Remember, if you attended the JALT 95 Conference, you do not need to order a copy.*

JALT Applied Materials

Language Testing in Japan

JAMES DEAN BROWN & SAYOKO OKADA YAMASHITA, EDITORS

DALE T. GRIFFEE, SERIES EDITOR

GENE VAN TROYER, SUPERVISING EDITOR

The Japan Association
for Language Teaching
全国語学教育学会
Tokyo, Japan

JALT Applied Materials

A series sponsored by the Publications Board of the Japan Association for Language Teaching.

JALT President: David McMurray
Publications Board Chair: Gene van Troyer
Cover design and graphics: Junji Maeda, Press Associates Nagoya
Design and typography: The Word Works, Ltd., Yokohama
Printing and binding: Koshinsha K.K., Osaka

LANGUAGE TESTING IN JAPAN

Copyright © 1995 by the Japan Association for Language Teaching, James Dean Brown and Sayoko Yamashita.

All rights reserved. Printed in Japan. No part of this book may be used or reproduced in any form whatsoever without written permission of the editors, except in cases of brief quotations embodied in scholarly articles and reviews. For information address JALT, 2-32-10-301 Nishi-Nippori, Arakawa-ku, Tokyo 116, Japan.

Cataloging Data

Brown, James Dean, & Yamashita, Sayoko Okada (eds.)
Language testing in Japan.

Bibliography: p.
1. Applied linguistics—Language testing—Second language learning. I. Title
1995
ISBN 4-9900370-0-6

Contents

Forward ..... *v*
Dale T. Griffee, Series Editor

About the Contributors 1

1 Introduction to *Language Testing in Japan* 5
James Dean Brown & Sayoko Okada Yamashita

SECTION I: CLASSROOM TESTING STRATEGIES

2. Differences between norm-referenced and criterion-referenced tests 12
James Dean Brown

3. Criterion-referenced test construction and evaluation 20
Dale T. Griffe

4. Behavioral learning objectives as an evaluation tool 29
Judith A. Johnson

SECTION II: PROGRAM-LEVEL TESTING STRATEGIES

5. Developing norm-referenced tests for program-level decision making 40
James Dean Brown

6. Monitoring student placement: A test-retest comparison 48
Sayoko Okada Yamashita

7. Evaluating young EFL learners: Problems and solutions 57
R. Michael Bostwick

SECTION III: STANDARDIZED TESTING

8. Good and bad uses of TOEIC by Japanese companies 66
Marshall Childs

9. A comparison of TOEFL and TOEIC 76
Susan Gilfert

10. English language entrance examinations at Japanese universities: 1993 and 1994 86
James Dean Brown & Sayoko Okada Yamashita

11. Exploiting washback from standardized tests 101
Shaun Gates

SECTION IV: ORAL PROFICIENCY TESTING

12. Testing oral ability: ILR and ACTFL oral proficiency interviews 108
Hiroto Nagata

13. The SPEAK test of oral proficiency: A case study of incoming freshmen 119
Shawn Clankie

14. Making speaking tests valid: Practical considerations in a classroom setting 126
Yuji Nakamura

SECTION V: INNOVATIVE TESTING

15. Cooperative assessment: Negotiating a spoken-English grading scheme with Japanese university students 136
Jeanette Mclean

16. Assessing the unsaid: The development of tests of nonverbal ability 149
Nicholas O. Jungheim

17. Cloze testing options for the classroom 166
Cecilia B. Ikeguchi

18. The validity of written pronunciation questions: Focus on phoneme discrimination .. 179
Shin'ichi Inoi

Foreword

About JALT Applied Materials

Welcome to the JALT Applied Materials (JAM) series, of which *Language Testing in Japan* is the inaugural (and pilot) project. The concept behind JAM is that each proposed volume is a collection of papers focused on a single theme, written by classroom teachers—some of whom are experienced writers and researchers and some who are just beginning. It is a concept that has been received with widespread, even enthusiastic approval from language teaching professionals in Japan, and we believe that it is sure to stimulate much fruitful discussion and continuing research in years to come. We are especially pleased to be able to bring this collection out in conjunction with JALT 95: the 21st International Conference on Language Teaching/Learning and Educational Materials Exposition.

The Situation

First, a few words about the reasons for embarking on this kind of publishing project. At present, books and collections of papers on various subjects are written and published by authors from various parts of the world, mainly from English-speaking countries such as Australia, Canada, New Zealand, the U.K. and the U.S.A. The authors of these books are typically university professors who teach courses in graduate programs and speak at conferences around the world, including JALT's annual international conference. There exists, however, a gap between the international level where these books are generated and written and the local teacher community in Japan and Asia. With rare exceptions, these books and collections—while of general professional value—bear little relevance to the L2 context in either Japan or Asia. We propose this new series of collected papers to help fill this gap, and JALT, as the largest L2 professional organization in Asia, is uniquely positioned to help bring this about.

Purpose

The JAM series is targeted at a) improving the quality of research and academic writing in Japan and Asia; b) publishing collections of articles on subjects of direct interest to classroom teachers which are theoretically grounded, reader-friendly and classroom-oriented; c) giving Japan- and Asia-based classroom teachers a publication outlet not heretofore available; and d) to help teachers around the world implement new ideas in their classrooms.

How JAM Works

We feel that JALT has reached a level of professional development as an organization in which many of its members are capable of writing creative and professional papers. Research and writing, however, are not easy tasks and we believe that more JALT members would be more capable of producing quality papers if they had comprehensive and constructive editorial support and guidance. To this end, the JAM project is guided by an editorial network composed of a general series editor, one or more country-based "local" editors and the possibility of an international editor for each collection of papers, all of whom would work with the authors offering contributions to a specific collection, not to mention the editorial expertise of JALT Publications' other editors.

Any person who is a resident of Japan, a member of JALT, and primarily a classroom teacher could request to be a JAM editor. Persons making such a request will be asked to submit a proposal stating the theme or area of focus for the collection, along with a CV and publications relevant to their subject. These Japan-based local editors could, if they desired, ask an international expert—for the sake of discussion, a "global" editor—in the field to assist them. After approval by the JAM committee and JALT Publications Board of the theme and editors, the next step would be for the local editor(s) to issue a call-for-abstracts. Deadlines for receipt of first and final drafts would be set by the local and series editors, and manuscripts must be approved by all editors: first by the local editor, then the global editor, and finally by the series editor. JAM collections are vetted through peer review, meaning that publication in one of them is a personal as well as professional achievement.

Contributing authors to a JAM collection are expected to be classroom teachers in Japan or a country with similar teaching conditions (e.g., Korea, China, Taiwan, Thailand, etc.), to be familiar with JALT and its mission, and to submit an original article which has not been published in or submitted to another publication. The JAM editors welcome contributions from new as well as experienced writers, and they need not have published previously in the theme area.

Conclusion: A Unique Opportunity

JALT is a professional organization of classroom teachers who occupy the niche between new developments and traditional classroom practices. As such we are historically positioned to creatively interact between theory, research and the daily reality of the language classroom. We firmly believe that what we have to report will be of significant value to teachers throughout Asia and around the world. We are confident that you will concur.

Dale T. Griffee
Series Editor

N.B.: Persons who desire further information about JAM should contact the current series editor, Dale T. Griffee, at: Koruteju #601, 1452 Oazasuna, Omiya-shi, Saitama-ken 330, Japan.

About the Contributors

Mike Bostwick has lived in Japan for over 10 years and is a doctoral candidate in the TESOL Program at Temple University Japan. He is also director of the English Immersion Program at Katoh Gakuen, a private Japanese school in Numazu, Shizuoka. Students in the program receive 50% or more of their regular school instruction in the English language on a daily basis. The program begins in preschool at three years of age and is expected to continue through high school when fully implemented. Throughout the program, Japanese children acquire English proficiency naturally as they study the regular concepts and skills required in the Japanese Ministry of Education curriculum.

James Dean ("JD") Brown, Professor on the graduate faculty of the Department of ESL at the University of Hawaii at Manoa, specializes in language testing, curriculum design, program evaluation, and research methods. He has taught in France, the People's Republic of China, Saudi Arabia, Japan, Brazil, and the United States. He has served on the editorial boards of the *TESOL Quarterly* and *JALT Journal*, as well as on the TOEFL Research Committee, TESOL Advisory Committee on Research, and Executive Board of TESOL. In addition to numerous journal articles and book chapters, he has published three books entitled: *Understanding Research in Second Language Learning: A teacher's guide to statistics and research design* (Cambridge, 1988); *The Elements of Language Curriculum: A systematic approach to program development* (Heinle & Heinle, 1995); and *Testing in Language Programs* (Prentice-Hall, 1995).

Marshall R. Childs is a lecturer at Fuji Phoenix College in Gotemba, Japan, and a program manager in the international department of Katoh Schools in Numazu, Japan. He is a consultant in intercultural and language education. Mr. Childs was a market research manager for IBM in the United States until 1985. He then worked for five years in IBM's Asia Pacific headquarters in Tokyo. Since 1990, he has been a teacher, curriculum writer, and consultant in Japan. Mr. Childs holds an A.B. from Harvard and an M.B.A. from New York University. He is currently a doctoral candidate in the TESOL Program at Temple University Japan. His research interests include psycholinguistic development and the integration of testing and curriculum.

Shawn M. Clankie is an American-born EFL instructor, originally from Rockton, Illinois. He holds a B.A. in French, and M.A. in EFL, from Southern Illinois University, and was also educated at l'Université de Savoie in France, and l'Université Catholique de Louvain-la-Neuve in Belgium. Teaching has taken him to the intensive ESL program at St. Johnsbury Academy in Vermont, USA, and then on to two years as a visiting lecturer at Kansai Gaidai University in Osaka. An avid writer, traveller, and language learner, his research interests include TESOL, historical linguistics, and Japanese linguistics. His articles have appeared in publications such as *The Language Teacher* and *TESL Reporter*, and his current research is on the internal and external factors of lexical change in pre-to-post World War II Japanese. In 1994, he moved to England to begin work towards a Ph.D. in Linguistics and is currently a M.Phil. candidate in theoretical Linguistics at the University of Cambridge, England.

Shaun Gates received his M.Sc. degree in Applied Linguistics from Edinburgh University. He is currently lecturing in English conversation and composition at Shiga Women's Junior College in Shiga Prefecture, Japan. He is also an oral examiner for the University of Cambridge Local Exam Syndicate (UCLES) and an oral and writing examiner for the International English Language Testing Service (IELTS).

Susan Gilfert received her M.A. degree from Ohio University and is currently lecturing at Nagoya University, Aichi Prefectural University, and Nagoya University of Foreign Studies. She also consults with Kawai Juku in Nagoya. She is co-author of *TOEIC Strategies*, a test-preparation text to be published by Macmillan Japan in fall 1995. She has written other texts including books on TOEFL preparation, study-abroad preparation, a global issues discussion text, and a film appreciation demonstration text for Kawai Juku. She has studied and presented papers on comparative standardized testing at JALT and TESOL conferences, and is interested in learning more about other standardized tests (such as the STEP series of tests) in Japan.

Dale T. Griffee, Assistant Professor at Seigakuin University, has taught in Japan since 1976 at commercial English language conversation schools, company in-house programs, junior colleges, intensive language programs, and universities. He graduated from Baylor University and holds an M.A. in TESL from the School for International Training, Brattleboro, Vermont. His research interests are SLA and the effects of instruction. He is especially interested in testing, interlanguage pragmatics, and the learner-centered classroom as an alternative to the traditional lecture-format, teacher-fronted classroom. He is author of three textbooks: *Listen and Act* (1982, Lingual House), *HearSay* (1986, Addison-Wesley), and *More HearSay* (1992, Addison-Wesley). His most recent publication is *Songs in Action* (1992, Prentice-Hall). He has held several offices in JALT including chapter president of the Sendai Chapter and the West Tokyo Chapter, as well as having served as member of the Long-Range Planning Committee, and Nominations and Elections Committee. He is general editor of the new JALT Applied Materials (JAM) series.

Cecilia B. Ikeguchi is a lecturer at the English Language Department at Dokkyo University and Tsukuba University in Japan. She obtained an M.A. degree in the teaching of English, and a Ph.D. degree in Education from the University of Tsukuba, a national university in Japan. She has been a teacher of English as a second and foreign language for ten years. Her research interests include ESL testing and bilingualism, as well as comparative education and educational systems.

Shin'ichi Inoi was born and brought up in Aizu-wakamatsu-shi, Fukushima Prefecture. He received his Bachelor of Education degree from Fukushima University in 1979 and taught English at several high schools in Chiba and Fukushima Prefectures for more than 10 years. He received a Masters of Education degree from Fukushima University in 1991 and has been a lecturer in the Literature Department of Ohu University in Koriyama since 1993, where he teaches phonetics, grammar, and linguistics. He is a member of JALT, JACET, and the Tohoku English Language Education Society. His research interests are in second language acquisition.

Judith Johnson, Associate Professor of Education at Kyushu Institute of Technology in Japan, has taught ESL/EFL, intercultural communication, linguistics, and teacher education courses as well as conducted teacher-training workshops in the Americas and Asia. Her interest areas are teacher training, curriculum design, materials development, and global education.

Nicholas O. Jungheim, Associate Professor at the Center for Languages and Cultural Exchange of Ryutsu Keizai University in Ibaraki Prefecture, is a graduate of the TESOL Program at Temple University Japan. His research interests include issues in classroom-oriented language acquisition, language learners' use of nonverbal communication, and language testing. He has presented his research on nonverbal communication at TESOL, AERA, and JALT conferences.

Jeanette McLean has been involved since the beginning of her career in various facets of education and has lived and worked in Europe, Africa, and Asia. With an educational background based in EFL, personnel management, and training, she has had the opportunity to teach at primary, secondary, and adult levels. Both in a training and management capacity, she has also been involved in and instigated development projects to assist and encourage educationally disadvantaged learners. As a result of this experience, she developed an interest in intercultural communication. Her subsequent research at Tokyo Denki University, where she taught for the last three years, focused on spoken English, the assessment of spoken English, and difficulties involved in oral interaction.

Hiroto Nagata has worked for over a decade as a language program coordinator at such major Japanese companies as Marubeni, Nissan, and Panasonic, where he has conducted a series of language proficiency interviews. He has published a number of articles on language teaching, and a variety of books and dictionaries for both children and adult learners of English. He currently teaches at Meikai University and Panasonic Tokyo, and is a doctoral candidate in the TESOL Program at Temple University Japan, where he also received his M.Ed. in TESOL.

Yuji Nakamura, Associate Professor of English at Tokyo Keizai University, teaches courses in language testing and EFL. He received his Ph.D. in Applied Linguistics (specializing in Language Testing) in 1994 from International Christian University in Tokyo. His research interests include large-scale assessment of the oral proficiency of Japanese students, application of item response theory to listening comprehension tests, and detailed study of validity.

Sayoko Okada Yamashita is currently teaching Japanese as a second language in the Division of Languages at International Christian University and the teaching practicum for the Labo International Foundation. She has also been developing audio-video materials for language teaching as a committee member of the National Language Research Institute. She has published a textbook for Japanese language teaching and articles on Japanese language teaching and Applied Linguistics. She received her Master of Education degree in Curriculum and Instruction at the University of Houston, Texas. She is now a doctoral candidate in the TESOL Program at Temple University Japan in Tokyo. Currently her primary research interest is in the testing of pragmatics.

Chapter 1

Introduction

JAMES DEAN BROWN
SAYOKO OKADA YAMASHITA

The idea for a book that focused on language testing issues in Japan took shape several years ago. It was originally suggested by Dale Griffee and then fleshed out as an idea and proposal by J. D. Brown (from his position at the University of Hawaii at Manoa in Honolulu). We knew that there were many books written about language testing as well as a number of edited collections of articles, but to our knowledge, no book had ever been dedicated specifically to the issues involved in language testing *in Japan*. The project became truly international when Sayoko Yamashita was invited to participate at logistical and editorial levels from her position at International Christian University in Tokyo.

Given that JALT was not in the habit of publishing books on professional issues in language teaching, considerable debate accompanied JALT's eventual approval of the idea of producing this book. Nevertheless, once the idea was given a green light, the production of *Language Testing in Japan* was taken up enthusiastically by the JALT. The call for papers went out through JALT's publications, *JALT Journal* and *The Language Teacher*, and a wide variety of interesting papers were received. In the fall of 1994, the papers in this book were selected, edited, and sent back to the authors for revisions. The contributors' final versions were submitted in winter 1995. Since then, the papers have been edited several more times in the process of producing this book.

From our perspective, the purpose of the book is to stimulate discussion about the testing that is done in the many language programs and classrooms throughout Japan. Naturally, testing serves many purposes in Japan including making decisions about students' language proficiency and placement, as well as for diagnosing their strengths and weaknesses, monitoring their progress, and measuring their achievement.

To help explore all of those uses of tests, the book offers chapters on classroom testing strategies, program-level testing strategies, standardized testing, oral proficiency testing, and innovative testing. To give readers an overview of the book, we will use the remainder of this chapter to summarize the other chapters. These summaries are based on the abstracts written by the authors themselves (please note that all citations will be found in the references of the respective chapters).

Classroom Testing Strategies

You are now reading Chapter 1, the introduction to the book. Next, you will find a section on classroom testing strategies, which includes Chapters 2, 3, and 4.

Chapter 2, entitled "Differences between norm-referenced and criterion-referenced tests" by James Dean Brown, addresses the distinction between norm-referenced tests (NRTs) and criterion-referenced tests (CRTs). He begins by examining differences in their test characteristics, including differences in: (a) the underlying testing purposes, (b) the types of decisions that are made with each test type, (c) the levels of generality of the material contained in the tests, (d) the students' expectations when they take the tests, (e) how scores are interpreted, and (f) how scores are reported to the students. Then, logistical differences between the two types of tests are discussed, including group size, ranges of ability, test length, time allocation, and cost. Finally, the author argues that: (a) the testing that teachers are doing in their classes is at least a good start, (b) scores on classroom CRTs may not necessarily be normally distributed, (c) certain testing responsibilities ought to rest with teachers and others with administrators, (d) CRTs and NRTs should be developed in different ways, and (e) NRTs cannot be expected to do all things well.

Chapter 3, entitled "Criterion-referenced test construction and evaluation" by Dale T. Griffee, introduces the subject of classroom test construction and evaluation to second language teachers, especially those teachers not familiar with test analysis procedures. The differences between CRTs and NRTs are touched on. The development and administration of pretests and posttests are explained. The subjects were 43 students in two classes at Seigakuin University. The chapter reports the results of those two test administrations. Descriptive statistics are presented and the Item Facility (IF) and Difference Index (DI) statistics are given for the first 15 items. Item analysis is discussed in some detail. The chapter concludes with a discussion of the implications of these classroom test construction procedures and the importance of institutional learning objectives.

Chapter 4, entitled "Behavioral learning objectives as an evaluation tool" by Judith A. Johnson, argues that one important key to enhancing student learning is to set and achieve meaningful and unambiguous instructional objectives. Stating objectives as specific, observable, and measurable learning outcomes that must be demonstrated by the student is essential to the instructional process: teaching/learning and evaluation. Bloom's taxonomy (1956) of observable student behaviors provides a practical and flexible framework within which (a) teachers can determine their instructional objectives, methods, and materials, and (b) teachers and students can measure student achievement during both the teaching/learning and evaluation phases of the instructional process, enabling them to determine which skills each student has mastered and which skills still need to be honed in order to fulfill the requirements of each objective. This chapter not only demonstrates how to write behavioral learning objectives, but also how to incorporate them into the instructional process as measurement tools in order to improve student learning.

Program-Level Testing Strategies

The next section of the book, consisting of Chapters 5, 6, and 7, groups three papers on program-level testing strategies.

Chapter 5, entitled "Developing norm-referenced language tests for program-level decision making" by James Dean Brown, describes the basic procedures involved in developing norm-referenced tests. Administrators and teachers can benefit in a number of ways from developing effective norm-referenced tests that can be used for aptitude, proficiency, and placement decisions. To help language professionals do so, this chapter addresses the following issues: (a) What are norm-referenced tests, and what are they used for? (b) What do administrators need in their norm-referenced tests? And, (c) how should norm-referenced tests be developed and improved?

This chapter also argues that norm-referenced tests must be accorded a much more important place in the administrator's program development plans. Strategies are suggested for finding or training staff members in this important area of curriculum development so that the important decisions that need to be made with norm-referenced tests can be made efficiently, effectively, and responsibly.

Chapter 6, "Monitoring student placement: A test-retest comparison" by Sayoko Okada Yamashita, investigates potential proficiency differences between subjects in the Fall and Spring terms of the Japanese Language Program (JLP) at International Christian University, along with possible reasons for such differences. The study is based on three subtest scores of the JLP placement test given in Fall as pretests and the same subtests given in Spring as posttests. Three research questions are addressed: (a) Do continuing students in the spring courses perform the same as those who had been placed into the equivalent courses in the Fall? (b) If there are differences in performance, are the differences observed at all levels or in some particular levels only? (c) Are there any differences in student performance on the different subtests? The study reveals that there are differences between the Fall and Spring populations. Possible reasons of this phenomenon are discussed and pedagogical implications are given.

Chapter 7, entitled "Evaluating young EFL learners: problems and solutions" by R. Michael Bostwick, argues that, although some types of paper and pencil tests may be appropriate for assessing language skills in older EFL students, they are not likely to be very suitable for younger EFL learners. Teachers who work with younger children need alternative assessment procedures that can inform instruction and provide teachers and students with feedback on the effectiveness of classroom teaching and learning. Before alternative forms of foreign language assessment can be employed, a number of issues must be considered so that the assessment procedures that are developed are aligned with the goals and objectives of the program, as well as with the actual instructional practices. This chapter examines several common problems foreign language programs for children face as they begin to consider ways to evaluate student progress and the effectiveness of their instruction. Solutions to these problems are discussed and two examples of assessment instruments are provided in order to help other programs begin to think about and develop their own assessment procedures.

Standardized Testing

This next section of the book includes Chapters 8 to 11. Chapter 8, by Marshall Childs and titled "Good and bad uses of TOEIC by Japanese companies," examines the *Test of English for International Communication* (TOEIC), which enjoys wide popularity among those Japanese companies that encourage their employees to learn English. One such company uses TOEIC scores to gauge the learning achievement of groups and to measure the effectiveness of different schools or treatments, and to guide individuals in choosing curricula. To what extent are these wise uses of TOEIC? A study of the scores of 113 learners in four successive tests during a course of English instruction in this company yields some answers. With proper care, TOEIC can be used to measure the learning gains of groups. For individuals, however, the variability of scores makes gauging learning uncertain, and the absence of achievement measures makes guiding learner curricula difficult. Advice is offered to administrators who may want to use TOEIC for gauging learning.

Chapter 9, "A comparison of TOEFL and TOEIC" by Susan Gilfert, begins by pointing out that the TOEFL is a well-known multiple-choice measure of an examinee's receptive English skills and is considered a reasonably good predictor of the examinee's productive skills. The general register of the TOEFL is academic English. The TOEIC is a lesser-known multiple-choice measure of an examinee's English skills in the general register of real-life, business situations. The TOEFL is created, produced, and sold by the Educational Testing Service (ETS) in Princeton, New Jersey. The TOEIC was originally created by ETS, but is now entirely owned and operated by the Japanese TOEIC office in Tokyo. This chapter gives a short history of the two tests, the TOEFL coming from American academic requests of the 1960s and the TOEIC coming from a Japanese Ministry of International Trade and Industry request in the middle 1970s. While both are multiple-choice, norm-referenced tests, the content of each test differs, as does the register (academic versus business English). Both tests measure English skills, but the purpose of the TOEIC is to provide "highly valid and reliable direct measures of real-life reading and listening skills"

(Woodford, 1982, p. 64). The chapter examines the skills tested by the TOEFL and the skills tested by TOEIC, and reports the apparent similarities and differences in both tests, as well differences in how they are scored and used.

Chapter 10, entitled "English language entrance examinations at Japanese universities: 1993 and 1994," by James Dean Brown and Sayoko Okada Yamashita, defines and investigates the various types of test items used in the English language entrance examinations administered at various Japanese universities in 1994 and compares these 1994 results to a previous study of the 1993 examinations. Ten examinations were selected from private universities and 10 from public universities along with the nationwide "center" examination. These twenty-one examinations were studied to find answers to the following questions: (a) How difficult are the various reading passages used in the 1994 university English language entrance examinations? (b) Are there differences in the levels of reading passage difficulty in private and public university examinations in 1993 and 1994? (c) What types of items are used on the 1994 English language entrance examinations, how varied are they, and how does test length vary? (d) Are there differences in the types of items and test lengths found in private and public university examinations in 1993 and 1994? (e) What skills are measured on the 1993 and 1994 English language entrance examinations?

To answer these questions, the examinations were analyzed item-by-item by a native speaker of English and a native speaker of Japanese. Then computer software was used to further analyze the level of difficulty of the reading and listening passages. The results should help English teachers in Japan to prepare their students for taking such tests and to help their students in deciding on which tests to take. Equally important, this study may continue to provide some professional pressure on those responsible for creating entrance examinations to do the best they can to create high quality tests.

Chapter 11 is "Exploiting washback from standardized tests," by Shaun Gates. It explores the washback effect, i.e., the influence of testing on teaching and learning. If teachers and students know the format and content of a test, they will tend to alter their classroom behavior to maximize performance on the test. Whether washback from any particular test is considered negative or positive depends on one's point of view. In this chapter, washback is considered positive if it promotes communicative language use in the classroom. Some of the factors identified as contributing to washback are particularly strong when considering standardized tests. Does standardized testing lead to a conflict with what the teacher wishes to achieve in the classroom? To answer this question, the written section of the IELTS test and the oral section in the UCLES Preliminary English Test (PET) are examined. The author argues that teachers preparing their students for these tests can exploit washback for classroom purposes.

Oral Proficiency Testing

The next section of the book, consisting of Chapters 12, 13, and 14, is about oral proficiency testing, which the Japanese educational system as a whole has only recently begun to consider.

In Chapter 12, "Testing oral ability: ILR and ACTFL oral proficiency interviews," Hiroto Nagata argues that, as language teachers, we are constantly facing the problem of reliably appraising our students' oral proficiency, and are always looking for some rule-of-thumb, easy-to-handle yardsticks. This chapter provides one such solution. Presented here are two interview models (the ILR and ACTFL oral proficiency interviews) with some practical suggestions on how to use them effectively not only for assessing students' oral proficiency but also for making decisions about (a) the kinds of remedial treatments that should be offered to students and (b) how to successfully add real-world reality to classroom activities as well.

Chapter 13, "The SPEAK test of oral proficiency: A case study of incoming freshmen" by Shawn M. Clankie, provides a detailed analysis of an attempt to institute the SPEAK test of oral proficiency, as a measure for assessing incoming freshmen, into an existing EFL program at Kansai Gaidai University in Osaka. The SPEAK test, based on retired TSE (*Test of Spoken English*) tests, contains six taped sections. Each student responds to the taped sections, and their voices are then recorded on separate cassettes for later evaluation. The test was administered to 150 students already admitted into the Intensive English Studies (IES) program on the basis of their TOEFL scores alone. This administration served as a preliminary run in preparation for candidates seeking admittance into the following year's classes. The TSE was administered and critiqued in terms of both its practicality and its ability to be accurately applied to Japanese learners. The chapter also describes the SPEAK test and the IES program, followed by a discussion of the benefits and failures of the TSE. In the end, the SPEAK test was abandoned in favor of more traditional face-to-face methods of oral evaluation. The reasons for this decision are explained.

Chapter 14, entitled "Making speaking tests valid: Practical considerations in a classroom setting" by Yuji Nakamura, argues that the central problem in foreign-language testing, as in all testing, is validity. Although there are other aspects such as reliability and practicality, there is no doubt that validity is the critical element in constructing foreign language tests. Validity concerns the question "How much of an individual's test performance is due to the language abilities we want to measure?" (Bachman, 1990). In other words, validity is about how well a test measures what it is supposed to measure (Thrasher, 1984). In this chapter, the author will (a) examine various kinds of validity in language testing in general, (b) discuss which types of validity are most suitable for a test of English speaking ability in a classroom setting, and (c) describe a validity case study which introduces the types of validity (i.e., construct validity, concurrent validity, face validity, and washback validity) which he has used in the development of a semi-direct speaking test suitable for use with Japanese college students. The construct validity of the test was examined through factor analysis and task correlation analysis; concurrent validity was estimated by the comparison of the test results and teacher's estimates; face validity, and washback validity were examined using a questionnaire and interviews.

Innovative Testing

The final section of the book, including Chapters 15 to 18, addresses the matter of innovative testing. In Chapter 15, "Cooperative assessment: Negotiating a spoken-English grading scheme with Japanese university students" Jeanette McLean describes how a workable grading scheme for assessing Japanese university freshmen's spoken English performance can be arrived at through negotiation with the students themselves. The setting up of this project is briefly described, and some problems inherent in assessing spoken English in Japanese universities are discussed. The steps taken to develop the grading scheme are described to demonstrate how student participation can be encouraged and how, through the consultation process, ownership of the assessment scheme by learners can be promoted. Benefits to the participants are then considered. The chapter concludes that this approach to testing can be used as an effective teaching tool to increase confidence, build trust, and harness cooperation among learners, teachers, and assessors and that the development of a grading scheme to which all participants contribute is a vital step towards achieving a valid test of spoken English in the educational context of Japan.

In Chapter 16, "Assessing the unsaid: The development of tests of nonverbal ability," Nicholas O. Jungheim, presents a rationale and a framework for testing nonverbal ability as a part of language learners' communicative competence and describes the creation of two tests of nonverbal ability using traditional test construction methods. The nonverbal ability framework, based on the Communicative Language Ability model (Bachman, 1988, 1990), consists of nonverbal textual ability, sociolinguistic ability, and strategic ability. The first test is the Gestest, a test of the nonverbal sociolinguistic ability to interpret North American gestures. The collection of baseline data, the application of traditional item analysis procedures, and the results of a pilot administration are described. The second test consists of the Nonverbal Ability Scales, a series of scales for rating language learners' use of head nods, gaze direction changes, and gestures in conversations. These scales involve nonverbal textual ability, including frequency and appropriateness of head nods and gaze direction changes, and nonverbal strategic ability, including the appropriateness and compensatory usage of gestures. Baseline data-collection procedures, the construction and interpretation of the scales, and the results of their application are described. The validity and reliability of both tests are discussed. The chapter concludes with a discussion of problems related to testing nonverbal ability and future prospects.

Chapter 17, by Cecilia B. Ikeguchi and titled "Cloze testing options for the classroom," points out that an apparent dichotomy exists between theoretical research on language testing and actual classroom testing. Specifically, the concept of *cloze* has often been confined within the walls of empirical research, *R*, which has often been considered distinctly different from practical language testing situations, labeled *r* by Lo Castro (1991). This chapter brings the results of research on cloze testing into the practical classroom setting by offering a simplified review of empirical studies on four major types of cloze tests and presenting practical classroom applications supported by the research. A chronological discussion of four major types of cloze tests (fixed-rate deletion, rational-deletion, multiple-choice cloze, and C-test) is presented along with their theoretical justifications, the features of each, and their distinct merits and demerits. The accuracy of measuring the specific language traits language teachers want to assess may depend on the kind of cloze procedure that is selected.

Chapter 18, "The validity of written pronunciation questions: Focus on phoneme discrimination" by Shin'ichi Inoi, examines the validity of phoneme discrimination questions for Center Exams administered by the Ministry of Education. To determine the validity of these types of questions, a written test of 30 phoneme discrimination questions and their oral versions were given to 60 Japanese college freshmen. One type of written phoneme discrimination question involves the learners being instructed to select from four alternative responses the word with the underlined part that is the same as the word they hear pronounced on a tape. The other type of written question requires the learners to choose the word with the underlined part pronounced differently from the other alternatives. On the oral version of the test, learners were asked to pronounce each of the words from the written test questions. The learners' pronunciations were recorded onto cassette tapes for later analysis. The data were analyzed in terms of the correlation of scores between the written and oral versions and the rate at which answers between the two tests agreed. The correlational analysis showed that moderately strong relationships existed between scores on the written oral tests, indicating, to some degree, the validity of the written phoneme discrimination questions. However, the agreement rate analysis also raised some doubts about their validity: low agreement rates prevailed, especially among the answers obtained from low-score achievers; on some items, the agreement rate failed to reach 50%.

Conclusion

To summarize briefly, ***Language Testing in Japan*** provides three to four chapters each on classroom testing, program-level testing strategies, standardized testing, oral proficiency testing, and innovative testing. We hope that readers will find this collection of papers on language testing in Japan as interesting as we do. More importantly, we hope that readers will find the papers accessible and useful, as well as provocative, and perhaps even inspiring.

Section I

Classroom Testing Strategies

Chapter 2

Differences Between Norm-Referenced and Criterion-Referenced Tests

JAMES DEAN BROWN
UNIVERSITY OF HAWAII AT MANOA

Language testing is an important aspect of the field of language teaching. Hence language teachers in Japan should be informed about the topic as should language teachers everywhere else. A number of books have been written on how to construct and use language tests (see, for instance, Alderson, Clapham, & Wall, 1995; Bachman, 1990; Baker, 1989; Brown, 1995a; Carroll & Hall, 1985; Harris, 1969; Heaton, 1988; Hughes, 1989; Madsen, 1983; or Valette, 1977), and many other books and articles have been written about theoretical aspects of language testing. Indeed, a journal called *Language Testing* has flourished for a number of years, and a conference occurs every year with the imposing name of Language Testing Research Colloquium. More importantly, in one way or another, language tests are used in almost every language program in the world in a variety of different ways including testing for: language aptitude, proficiency, placement, diagnosis, progress, and achievement.

With all these important uses, tests ought to be of interest to almost every language teacher. Yet in my professional travels, especially in Japan, I have found that most teachers are intimidated by tests and maybe even a little frightened of them. Why is that? I believe that one cause is the fact that nobody has ever taken the trouble to clearly explain to most teachers how very useful tests can be to them in their daily work. Hopefully, many of the articles in this book will provide explanations and examples of just how useful tests can be for language teachers in Japan.

As a basis for that process, I will begin in this chapter by explaining the differences between criterion-referenced and norm-referenced language tests. Some teachers might argue that such jargon (norm-referenced and criterion-referenced) is one of the primary problems with language testing. Perhaps that is true, but nonetheless, the jargon exists, and I think that this particular jargon is very important because understanding these particular concepts is crucial if language teachers are to truly understand the variety of useful ways that tests can function in their program decision making and in their individual classes.

I'll address the distinction by answering two central questions: How are criterion-referenced tests fundamentally different from norm-referenced tests? And, why is the distinction important to both language teachers and language program administrators?

Criterion-Referenced vs. Norm-Referenced

Criterion-referenced tests (CRT) and norm-referenced tests (NRT) differ in many ways, and an ever-increasing literature is developing within language testing on the distinction between CRTs and NRTs (Cartier, 1968; Cziko, 1982, 1983; Hudson & Lynch, 1984; Brown, 1984, 1988, 1989, 1990a & b, 1992, 1993, and 1995a & b; and Bachman, 1989, 1990). I have found that the primary differences between CRTs and NRTs can be classified as either differences in test characteristics or differences in the logistics involved. Hence, the following discussion will be organized around those two basic categories.

Table 1. *Differences in Characteristics and Logistics for CRTs and NRTs (adapted from Brown, 1992)*

	CRTs	NRTs
Test Characteristics		
Underlying Purposes	foster learning	classify/group students
Types of Decisions	diagnosis progress achievement	aptitude proficiency placement
Levels of Generality	classroom specific	overall global
Students' Expectations	know content to expect	do not know content
Score Interpretations	percent	percentile
Score Report Strategies	tests and answers to	only scores go to students
Logistical Dimensions		
Group Size	relatively small group	large group
Range of Abilities	relatively homogeneous	wide range of abilities
Test Length	relatively few questions	large number of questions
Time Allocated	relatively short time	long (2-4 hrs) administration
Cost	teacher time & duplication	test booklets, tapes, proctor & a fee

Differences in Test Characteristics

As I have pointed out elsewhere (Brown, 1990b, 1992, & 1995a), six primary test characteristics distinguish criterion-referenced from norm-referenced tests: differences in underlying purpose, in the types of decisions that are made, in the levels of generality, in students' expectations, in score interpretations, and in score report strategies. The upper portion of Table 1 summarizes how CRTs and NRTs differ on these important test characteristics.

Underlying purpose. The basic purpose of criterion-referenced tests is to foster learning. Typically, teachers administer CRTs in order to encourage students to study, review, or practice the material being covered in a course and/or in order to give the students feedback on how well they have learned the material. In contrast, the underlying purpose for norm-referenced tests is usually to spread students' performances out along a continuum of scores so the students can be classified or grouped for admissions or placement purposes. While creating such groupings may be beneficial to learning, the NRTs are not typically designed to test material that is specifically and directly related to a single course or program. Thus NRTs are not directly created to foster learning.

Types of decisions. Some types of testing decisions are best made with criterion-referenced tests. For instance, CRTs are well suited to making diagnostic, progress, and achievement decisions. Such decisions usually involve giving students feedback on their learning: either detailed feedback on an objective-by-objective basis in diagnostic and progress testing, or less detailed feedback in the form of course grades, which are very often based, at least in part, on achievement test scores. Norm-referenced tests are more appropriately used for aptitude, proficiency, and placement decisions. Aptitude and proficiency test scores are typically used for determining who will be admitted to a particular institution or program of study. Later, after students have been admitted to an institution, a placement test may be used for deciding the appropriate level of study within that institution.

Levels of generality. Criterion-referenced tests are usually based on the very specific objectives of a course or program. CRTs are therefore likely to vary in content, form, and function from teacher to teacher and from course to course. In contrast, norm-referenced tests are usually designed to be institution-independent, that is, NRTs are typically used with students from many institutions, or at least many classrooms. Norm-referenced tests must therefore be based on knowledge, skills, or abilities that are common to a number of institutions, programs, or courses.

Students' expectations. Students generally know what to expect on a criterion-referenced test. Indeed, they will ask the teacher what they should study for the test, and the teacher usually ends up playing a kind of game with them: giving a list of the points that have been covered in the course (that is, the course objectives that the teacher wants the students to study, review, or practice) without giving away exactly what will be on the test. But in any case, the students should know from the course objectives what will be covered on a CRT. On norm-referenced tests, the students usually have little idea of what content to expect. They may have some idea of the types of test questions that will be on the test (for instance, multiple-choice, interview, cloze passage, writing task, role-play, etc.), but they will have virtually no idea of the exact content that the test questions will cover. Security is often a very important issue on NRTs.

Score interpretations. Criterion-referenced tests are typically scored in percentage terms. For example, a student who scores 95 percent on a CRT might get an A grade, or 88 percent might earn a student a B+, or 33 percent might be failing. In contrast, norm-referenced test scores are typically reported as "standardized" scores like the CEEB scores reported for the overall *Test of English as a Foreign Language* (TOEFL) (ETS, 1995), where 500 is the norm average and the scores range from 200 to 677 (see Brown, 1995a for a full explanation of how these standardized scores work). Some NRTs report percentile scores, which are also standardized scores. Percentile scores are generally easier for students to understand. Such NRT scores only have meaning with reference to a specific population of students (i.e., the group of students which served as the norm group). Thus, on an NRT, a student who is in the 91st percentile has scored better than 91 out of 100 of the students in the norm group, but worse than 9 out of 100. On NRTs, the concern is not with how many questions the student answered correctly (or what percent), but rather with how the student scored relative to all of the other students in the norm group (that is, the percentile). Such interpretations are very different from the straightforward percent typically used in thinking about CRT performance.

Score report strategies. In order to foster learning through criterion-referenced testing, teachers will often return the tests to the students when reporting their scores, and even go over the answers in class so the students can benefit from the experience. Hence the students usually know exactly how many questions they answered correctly and which ones. On norm-referenced tests, students typically do not get their tests back after taking the examination.[1] In most cases, they are not even told the actual number of questions they answered correctly. Instead, they are given transformations of their actual "raw" scores into "standardized" scores. As a result, students seldom know how many questions they answered correctly on an NRT.

Logistical Differences

As shown in the bottom portion of Table 1, the logistical differences between CRTs and NRTs have to do with differences in group sizes, ranges of abilities, test length, time allocated, and cost.

Group size. Criterion-referenced tests are most often constructed to be used in classroom settings. Thus the group size involved is usually limited to the relatively small (say, 10 to 75 students) groups found in language classrooms. NRTs, on the other hand, are typically constructed to be administered to relatively large groups of students. For instance, the TOEFL has been administered to about a million students worldwide in each of the past three years (plus or minus a couple of hundred thousand). This fact, alone, has certain implications in terms of the test question formats that can be employed, and may account for the general prevalence of multiple-choice questions in norm-referenced tests. After all, multiple-choice questions are relatively easy to score, especially if the scoring can be done by machine.

Range of abilities. In addition, criterion-referenced tests are usually designed for a group of students who have been placed into a particular level of study and are studying exactly the same material. Thus the range of abilities involved in criterion-referenced testing situations is usually relatively narrow. In contrast, norm-referenced tests are typically normed on a population of students with very wide ranges of abilities. For example, the TOEFL is normed on a group with abilities that range from virtually no knowledge of the English language (that is, a student whose score is based entirely on guessing) to high native-speaker ability. Thus the TOEFL scores represent a very wide range of abilities, indeed.

Test length. Furthermore, CRTs tend to contain relatively few test questions because they are designed to test a relatively small body of knowledge or set of skills. In comparison, NRTs tend to be fairly long because of the large body of knowledge or skills that is being assessed. In addition, NRT developers are well aware that making their tests long enhances their chances of achieving good statistical characteristics—particularly reliability.

Time allocated. Because criterion-referenced tests are relatively short and are often administered during class time, they tend to be relatively quick, usually about one hour, depending on how long the class is scheduled to meet. Because norm-referenced tests have more questions and are not limited by class scheduling, they tend to take much more time to administer. The NRTs that I have worked on have generally ranged from about two hours to six hours.

Cost. The last logistical distinction between CRTs and NRTs involves cost. Criterion-referenced tests are usually viewed as being nearly free because they only involve the teacher's time (testing is usually considered part of the teacher's job) and whatever minimal duplication costs might be involved. Because norm-referenced tests are longer and more elaborate (including test booklets, tapes, answer sheets, official proctors, and considerable space), they usually cost much more to administer. Naturally, some NRTs are cheaper than others. For instance, the TOEFL with its multiple-choice machine-scorable formats is much cheaper than its companion tests, the *Test of Spoken English* (TSE) and the *Test of Written English* (TWE), which are also published by Educational Testing Service on a regular basis. The TSE and TWE are more expensive because they involve speech samples and written compositions, respectively. These productive language samples must each be judged by several paid human raters, and, of course, that is a relatively expensive process.

Two Contrasting Example Tests

In terms of the six test characteristics listed in Table 1, a typical example of a criterion-referenced test would be the intermediate level speaking test that we developed when I was teaching EFL in China from 1980 to 1982 (see Brown, 1995b, for further description of this program). The underlying purposes of this test were to foster learning by giving the students diagnostic feedback at the beginning of the course and progress feedback at the midterm, and by assessing their overall achievement at the end of the course. One of our hidden purposes was to get students who were accustomed to traditional grammar/translation language teaching to cooperate in role-play, pair-work, and other communicative activities that they thought were strange and pointless. Thus the content of the test was very specific and based on the precise objectives of our course. Our objectives were for the students to be able to effectively use 15 of the functions covered in the Gambits series (Keller & Warner, 1979) by the end of the course. It was very easy for the students to predict which functions were

BEST COPY AVAILABLE

going to be tested because we told them precisely what our course objectives were and precisely what to expect on the test in terms of objectives, directions, format, timing, and tasks. In addition, they took the same test three times so, at least by the end of the course, everyone was sure to know what to expect on the test. The test took the form of a taped interview. To save time, each student randomly selected three cards from the fifteen that we had (one for each objective), and the interview proceeded. Various schemes were used to score these interviews, but the best one from my point of view asked two teachers (the student's own teacher and one other) to give the students separate ratings for fluency, content, effectiveness at getting their meaning across, using the correct exponents to accomplish the function, and stress/intonation. Each of these five categories was worth 20 points for a total of 100 points, which we interpreted in percentage terms as is typical of a CRT. We didn't give the tests back to the students, but we did give them feedback (on the diagnostic and progress tests) in the form of a teacher conference with each student, where we played the tape for them and gave them oral feedback from our notes.

In terms of logistical dimensions, this test was also clearly a CRT. Every ten-week term, the test was typically administered three times (weeks one, five, and ten) to three small classes of 20 students. All of the students were in the intermediate level so they were fairly homogeneous in terms of their range of abilities. The test was also relatively short with only three questions per student in an interview that took only about five minutes; and, this speaking test cost very little to administer because the administration and scoring were considered part of the teachers' jobs, and the materials were duplicated as a normal part of program expenses.

An example of a norm-referenced test is the TOEFL (ETS, 1995), which can also be examined in terms of the six test characteristics listed in Table 1. The underlying purpose of the TOEFL is to classify students along a continuum of general abilities (that is, overall English language proficiency for academic purposes) usually with the ultimate goal of making admissions decisions at universities in the United States. Thus the content of the test is very general within three broad categories: (a) listening comprehension, (b) writing and analysis, and (c) vocabulary and reading comprehension. Within those categories, it is very difficult to predict exactly what the content of the questions will be, and Educational Testing Service strives to maintain that situation by using a large test question pool, by developing 12 forms per year, by taking legal action against those who violate their copyright, and by enforcing strict test security measures worldwide. Hence, it is difficult for students to know exactly what to expect on the test when they take it. The scores on TOEFL are always interpreted as standardized scores. For instance, a student who scores 500 is average in comparison to the norm group, that is, 50 percent of the students would be below the student in the distribution of scores (and of course, 50 percent would be above the student, too). In addition, Educational Testing Service would prefer not to give the tests back to students. Indeed, students have to go out of their way to request a copy of the test before ETS will supply it. This is reasonable from the tester's point of view because giving out copies of the tests compromises test security and raises the costs of testing—costs that will have to be passed on to future examinees.

In terms of logistical dimensions, the TOEFL is also clearly an NRT. Every year, it is administered to many hundreds of thousands of students around the world who have a wide range of abilities from virtually no English to high native proficiency. This test is also relatively long, both in the number of test questions and the time that it takes to administer it (about two hours and 30 minutes). Finally, the TOEFL costs a great deal more than a CRT to administer, and as a result, it has a registration fee that amounts to a month's salary in some countries.

In Japan, numerous examinations are, or should be, considered NRTs. For instance, because of the types of admissions decisions being made, the high school and university entrance examinations should probably be considered norm-referenced. Similarly, the TOEIC and the various Eiken examinations are norm-referenced.

Typically, norm-referenced tests in the United States are analyzed qualitatively and statistically to determine the degree to which they are functioning well as norm-referenced tests (particularly in terms of test-item quality, descriptive statistics, test reliability, and validity). However, after considerable effort, I have found it impossible to obtain similar information about the norm-referenced tests in Japan. This raises the question of whether the quality of NRTs in Japan is analyzed at all (statistically or otherwise).

Why is the Distinction Important?

The distinction between CRTs and NRTs is primarily important to teachers in Japan because, based on the distinction, they should realize that: (a) the testing that they are doing in their classes is at least a good start, (b) scores on a classroom test may not necessarily be normally distributed because such tests should be criterion-referenced, (c) certain testing responsibilities ought to rest with teachers and others with administrators, (d) CRTs and NRTs are developed in different ways, and (e) NRTs cannot be expected to do all things well.

Most current classroom testing is a good start. I'm sure that in many cases, classroom criterion-referenced testing could be done better, but at the same time, I am also sure that most teachers are at least on the right track when they try to test the things that they have taught in their courses. The only teachers who might be on the wrong track are those who don't do any testing at all. Students need feedback on how they are doing. If asked, most students will even say that they like to be tested. But, to be effective, a CRT must match what is being taught in the class. So it would be inappropriate to teach communicative language skills and functions and then test the students with a multiple-choice grammar test.

CRT scores may not be normally distributed. Teachers will also be comforted to recognize that a normal distribution (commonly known as the bell curve) may not necessarily occur in the scores of their classroom tests. As mentioned above the groups of students are usually small and homogeneous, and it is not reasonable to expect a normal distribution of scores in such groups. In addition, on CRTs, the ideal distributions would occur if all of the students scored zero at the beginning of a course (indicating that they all desperately needed to learn the material) and 100 percent at the end of the course (indicating that all of the students have learned all of the material perfectly). Neither of these ideals is ever really met, even with a good test, but the scores might logically be "scrunched up" toward the bottom of the range at the beginning of a course and toward the top of the range at the end of the course. Hence for a number of reasons, expecting a normal distribution in classroom testing is unreasonable. Nonetheless some administrators expect just that, usually in the name of "grading on a curve." Teachers now have the information to show administrators the error of their ways—though in many situations in Japan, teachers may prefer to keep their own counsel.

Testing roles of teachers and administrators. However, even in Japan, both teachers and administrators can benefit from thinking about the relationships of these different types of tests to their respective job responsibilities. I think that CRTs for testing diagnosis, progress, and achievement ought properly to be the responsibility of teachers, individually, or better yet, collectively in groups of relevant teachers working together. In contrast, any NRTs for testing aptitude, proficiency, or placement should primarily be the responsibility of the administrators. I am not saying that I think that administrators should not to be involved in the criterion-referenced testing processes or that teachers should not help with the norm-referenced testing administrations. In most cases, the administrators probably need to help the teachers coordinate the development of CRTs and help them with the logistics of administering and scoring them. Equally important, the administrators will probably need help from the teachers in proctoring and scoring a placement test, and maybe in developing it. Nonetheless, because of the nature of the respective decisions that will be made with these tests, I strongly feel that the primary responsibility for CRTs ought to rest with teachers, while the primary responsibility for the NRTs should rest with administrators.

CRTs and NRTs are developed in different ways. One other difference exists between CRTs and NRTs that is not covered in this chapter. It is the difference in the strategies that are used for developing and improving CRTs and NRTs. Those issues are covered in other chapters later in the book: one on CRTs and one devoted to NRTs.

NRTs can't do it all. Teachers and administrators should not expect too much of NRTs. Often in Japan, I have found that NRTs like the TOEFL are used for many types of testing, including proficiency, placement, achievement, and even progress testing. While the TOEFL, as a very general NRT, can serve as an excellent proficiency test, it seems irresponsible to use it as a placement test because, in most cases, the TOEFL is too broad in nature to work well in placing the students in a particular institution. Similarly, the content of the TOEFL is entirely too broadly defined to be useful in tracking the progress of students, or measuring their achievement in semester-length, or even year-long English courses.

As Alderson (1990) put it at the RELC Seminar on Language Testing and Language Programme Evaluation in Singapore, norm-referenced tests simply are not sensitive enough for doing criterion-referenced testing. This lack of sensitivity is certainly related to the degree of specificity involved, but may also be due in part to the ways that criterion-referenced and norm-referenced tests differ along logistical dimensions.

Administrators and teachers alike should also realize that using NRTs for CRT purposes minimizes the possibilities that their program will look good. Program evaluations conducted by smart language professionals will use tests that are sensitive to the goals and objectives of the program involved in order to maximize the chances of their students showing gains in learning. Such program-sensitive tests (sometimes called program-fair tests) are by definition criterion-referenced, not norm-referenced.

Conclusion

In sum, because of all of the differences listed in Table 1 and the forgoing discussion, it seems clear that the characteristics of criterion-referenced tests make them unsuitable for classifying students into groups for admissions or placement decisions, while at the same time, the characteristics of norm-referenced tests make them inappropriate for assessing what percent of the material covered in class each student has learned for diagnostic, progress, or achievement decisions. Hence, one of the few truths available to us in language teaching is that criterion-referenced and norm-referenced language tests are fundamentally different from each other.

Perhaps the single most important message that this chapter contains is that tests, whether criterion-referenced or norm-referenced, are important tools in language programs. These tools that can and should be used effectively and efficiently to help language teachers and administrators with the variety of different types of decisions that they must make in order to deliver excellent instruction to their students. However, all of this can only be accomplished if language professionals in Japan understand how to use both criterion-referenced and norm-referenced tests properly.

Note

[1] Note that the Educational Testing Service has recently developed a policy whereby students who take the TOEFL and want a copy of the examination can obtain it by writing directly to Educational Testing Service.

References

Alderson, C. (1990). Language testing in the 1990's: How far have we come? How much further must we go? Plenary address at the 1990 RELC Seminar on Language Testing and Language Programme Evaluation. Singapore.

Alderson, C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Bachman, L. F. (1989). The development and use of criterion-referenced tests of language proficiency in langauge program evaluation. In K. Johnson (Ed.) *Program design and evaluation in language teaching.* Cambridge: Cambridge University Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Baker, D. (1989). *Language testing.* London: Edward Arnold.

Brown, J. D. (1981). Newly placed versus continuing students: comparing proficiency. In J.C. Fisher, M.A. Clarke & J. Schachter (Eds.) *On TESOL '80 building bridges: Research and practice in teaching English as a second language* (pp. 111-119). Washington, DC: TESOL.

Brown, J. D. (1984). Criterion-referenced language tests: What, how, and why? *Gulf Area TESOL Bi-annual, 1,* 32-34.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design.* Cambridge: Cambridge University Press.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23*(1), 65-83.

Brown, J. D. (1990a). Short-cut estimates of criterion-referenced test consistency. *Language Testing, 7*(1), 77-97.

Brown, J. D. (1990b). Where do tests fit into language programs? *JALT Journal, 12*(1), 121-140.

Brown, J. D. (1992). Classroom-centered language testing. *TESOL Journal, 1*(4), 12-15.

Brown, J. D. (1993). A comprehensive criterion-referenced language testing project. In D. Douglas and C. Chapelle (Eds.) *A New Decade of Language Testing Research* (pp. 163-184). Washington, DC: TESOL.

Brown, J. D. (1995a). *Testing in Language Programs.* Englewood Cliffs, NJ: Prentice-Hall Publishers.

Brown, J. D. (1995b). *The Elements of Language Curriculum: A systematic approach to program development.* New York: Heinle & Heinle Publishers.

Carroll, B. J., & Hall, P. J. (1985). *Make your own language tests: A practical guide to writing language performance tests.* Oxford: Pergamon.

Cartier, F. A. (1968). Criterion-referenced testing of language skills. *TESOL Quarterly, 2*(1), 27-32.

Cziko, G. A. (1982). Improving the psychometric, criterion-referenced and practical qualities of integrative language tests. *TESOL Quarterly, 16*(3), 367-379.

Cziko, G. A. (1983). Psychometric and edumetric approaches to language testing. In J. W. Oller, Jr. (Ed.), *Issues in language testing research.* Cambridge, MA: Newbury House.

ETS. (1995). *Test of English as a foreign language.* Princeton, NJ: Educational Testing Service.

Harris, D. P. (1969). *Testing English as a second language.* New York: McGraw-Hill.

Heaton, J. (1989). *Writing English language tests (New edition).* London: Longman.

Hudson, T., & Lynch, B. (1984). A criterion-referenced approach to ESL achievement testing. *Language Testing, 1*(2), 171-201.

Hughes, A. (1988). *Testing for language teachers.* Cambridge: Cambridge University Press.

Keller, E., & Warner, S. (1979). *Gambits conversational tools* (Books one, two, & three). Hull, Quebec: Canadian Government Printing Office.

Madsen, H. (1983). *Techniques in testing.* Oxford: Oxford University Press.

Valette, R.M. (1977). *Modern language testing* (2nd ed.). New York: Harcourt Brace Jovanovich.

Chapter 3

Criterion-Referenced Test Construction and Evaluation

DALE T. GRIFFEE
SEIGAKUIN UNIVERSITY

Despite the increasing number of teachers who have master's degrees in teaching English as a second language (TESOL) or other forms of training, there remains an ignorance and even an aversion to the technical aspects of test construction on the part of many teachers. In the past three years, I have attempted classroom tests by several means, all of which proved unsatisfactory. Paper tests were either too easy or too difficult for my classes and interview tests proved exhausting for me. Eliminating tests altogether and basing grades on class participation and attendance also proved unsatisfactory for several reasons. First, I had the feeling of not being fair to my students. I flunked one student who was on the borderline of allowable absences, seldom participated in discussions, and on occasion slept in class. On the other hand, I gave high grades to students with good attendance but low class participation. In both cases, I felt on shaky ground and wished for additional criteria. Second, students expect a test, and I wonder how seriously they take a course without a final examination. Third, by not giving tests, I was not receiving any feedback on student progress. Fourth, without a pretest, I had no idea what the levels of my students were on entering my class or what their level of previous knowledge was. Fifth, without tests, especially a final test, I was getting no sense of closure and completion to my course. The purpose of this paper is to introduce criterion-referenced tests (CRT) to teachers who either dislike the whole idea of testing or who have for one reason or another avoided the issue of testing. Two questions will be addressed: What is the difference between criterion-referenced tests and norm-referenced tests (NRT)? And, how can criterion-referenced tests be evaluated and revised?

Norm-Referenced Tests and Criterion-Referenced Tests Defined

Most classroom teachers are familiar with norm-referenced tests by function if not by name. One of the well-known NRTs is the *Test of English as a Foreign Language* (TOEFL). However, few ESL/EFL teachers are familiar with the concept of criterion-referenced tests, perhaps because CRTs have not been discussed as much in the literature as NRTs. For example, in explaining NRTs and CRTs, one popular teacher training text (Savignon, 1983, p. 240) gives four paragraphs consisting of fifty-one lines and two tables to explaining NRTs, but gives only one paragraph consisting of six lines to explaining CRTs. Perhaps unfamiliarity with CRTs is due to the fact that NRTs have dominated testing since the mid-1970's (Bachman, 1989, p. 248). Perhaps another reason NRTs are more familiar to teachers than CRTs is because NRTs are

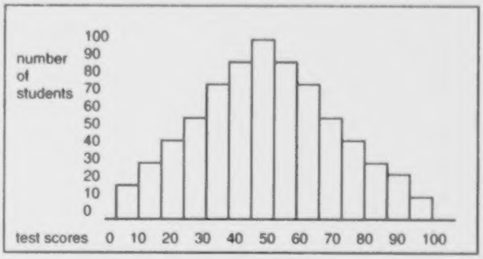
Table 1. *Differences between norm-referenced and criterion-referenced tests Adapted from Brown (1989).*

Characteristic	Norm-referenced	Criterion-referenced
Purpose of testing	To spread students out along a continuum of general ability	To determine the amount of material learned
Type of measurement	General language abilities are measured	Specific language points are measured
Type of Interpretation	Relative: A student's performance is compared with that of all other students	Absolute: performance is compared only with a pre-specified learning objective
Knowledge of questions	Students have little or no idea of what content to expect in the test questions	Students know exactly what content to expect in the test questions
Score distribution	Normal distribution of scores, e.g. a bell curve	If all students know all the material, all should score 100%

used to decide proficiency and placement issues which are of high interest to both program administrators and classroom teachers. For whatever reason, the distinction between NRTs and CRTs is only recently being recognized by TESOL teachers (Brindley, 1989, p. 49; Brown, 1990a, p. 125; Brown, 1992). Table 1 summarizes the differences between NRTs and CRTs.

NRTs measure general language proficiency whereas CRTs measure specific objectives (Brown and Pennington, 1991, p. 7). An NRT cannot give specific information relative to what a student can or cannot do with language, and it is this characteristic that leads Bachman (1989, p. 243) to say that NRTs are a poor choice for program evaluation. NRTs interpret student scores relative to other student scores, whereas on CRTs students' scores are interpreted relative to an absolute standard, e.g., learning 25 vocabulary words by the end of the week. For NRTs to successfully compare students, they must involve a large enough sample of students (30 or more) to create what is called a normal distribution (see Richards, Platt, and Platt, 1992, p. 249). Figure 1 shows an example of a normal distribution.

Figure 1. *Normal distribution*



If you were to connect the top of each bar with a line, you would see the familiar bell curve shape. A CRT, on the other hand, does not operate on the normal distribution concept. In fact, a good CRT would have a positively skewed distribution for the pretest, as seen in Figure 2, and a negatively skewed distribution for the posttest as seen in Figure 3.

Figure 2. *Positively skewed distribution*

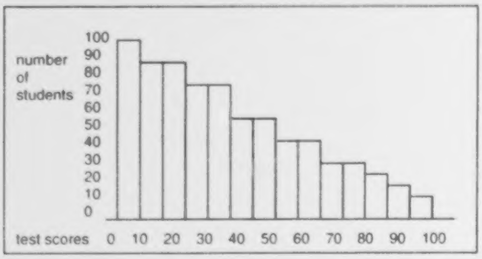
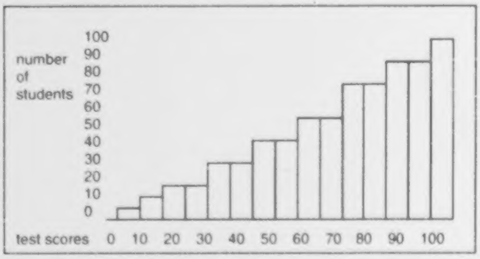


Figure 3. *Negatively skewed distribution*



The reason a well-functioning CRT pretest might produce the distribution in Figure 2 is that at the beginning of a course it might reasonably produce many scores at the low end because students did not have the knowledge or skill being tested. In other words, in an ideal situation a CRT pretest would indicate that many students did not know the material and scored zero or close to it. At the end of the course, when the students had the benefit of instruction and took the posttest, they would probably score very high, which would result in a distribution like that shown in Figure 3.

Method

Subjects

The subjects in this study were 50 second-year students at Seigakuin University in the newly-formed Division of Euro-American Studies. The students were in two classes, one of which met on Tuesday and another which met on Thursday. Each class met once a week for 90 minutes. The Tuesday class had 25 students consisting of 11 women and 14 men, and the Thursday class had 25 students consisting of 12 women and 13 men. The students ranged in age from 19 to 21 years of age. Due to absences and late registration, 43 students (21 women and 22 men) took the pretest and 37 students (20 women and 17 men) took the posttest. All students were Japanese and all except two were from Saitama prefecture and the nearby Tokyo area. No university or department objectives existed at the time of this study and no TOEFL or other NRT test scores were available. The course syllabus was entirely up to the instructor. One grade was to be given to each student for the entire year based on attendance, homework, and the final test.

Materials

In light of the absence of any institutional course objectives, the test was based entirely on the course textbook *More HearSay* (Griffee, 1992). Each unit in the textbook has approximately an equal number of listening and speaking exercises. The general criterion for test construction was that the test reflect the course book as much as possible. Other more specific criteria were that the test be a written test, at least half listening; that there be fifty items scored two points each; that the test contain no material directly taken from units one and two; that all questions be scorable as right or wrong; that all content items be explicitly taught in the text; and that, except for units one and two, the test items cover as much of the text as possible. Eleven formats used in the textbook were identified, and five were judged acceptable for inclusion in the test: cloze passages, multiple-choice questions, listen and write the word/number/prices you hear, listen and circle the word you hear, listen and identify what is described, and write the phrase you hear. Fourteen content areas were identified and eight were used because they had a wide distribution throughout the textbook: vocabulary, cultural items, numbers, schedules, cities, money, food, and travel. The final form of the test, not given here for test security reasons, had nine sections with a total of fifty items. The first six sections of the text, consisting of 26 questions, were for listening and included the following subtests: listen and write what you hear, count the number of words you hear in these sentences, listen and identify which state in the U.S. is being described. The last three sections contained only written items, which included matching words and specific content questions such as "What does ASAP mean?"

Procedures

The pretest was administered in April 1993 during the second class meeting of both classes. In both classes, the first meeting was taken up with a course introduction and Unit 1 of the textbook. No pretest makeup was administered to any student who was absent or transferred into the course after the second meeting. The posttest was administered in January 1994 during the last class session. Both the pretest and the posttest were administered using the same cassette tape, which included instructions as well as the listening passages. The tests were then collected and graded by the instructor, but not returned to the students. The statistics were calculated on a Macintosh LC 520 using the Claris Works version 2.0 spreadsheet program.

Analysis

In this paper three types of statistics will be discussed: descriptive statistics, item statistics, and consistency estimates.

Descriptive Statistics

Descriptive statistics, as the name implies, give a basic description of the test. In this paper, seven descriptive statistics will be given. Five of the statistics are self-explanatory: they are the number of students, the number of test items, the minimum score, the maximum score, and the range of scores. The remaining two statistics, the mean (which is sometimes symbolized by the letter M or $\bar{x}$ [pronounced ex-bar]) and standard deviation (SD), require some explanation. According to Richards, Platt, and Platt (1992, p. 349), the mean is the average of a set of scores. In other words, the mean is the sum of scores divided by the number of test scores. If the scores on a certain test are 2, 4, 6, and 4 their total is 16. Divide this sum by four (the number of scores) and the mean is 4. The standard deviation is an average of the difference of each score from the mean. The word "deviation" refers to how far each score is from the mean and the word "standard" is a kind of average. The formula for the standard deviation is as follows:

$$SD = \sqrt{\frac{\Sigma (x - \bar{x})^2}{N}}$$

Where: SD = standard deviation
$\bar{x}$ = mean
x = scores
N = number of students
S = sum

Item statistics

In a language test, each scorable piece of language is called an item, and a test question may contain one or more items. Item analysis is a way of obtaining some simple statistics to analyze the items on the test. Item analysis might be a new concept for many teachers, but it should be interesting for classroom teachers working with tests because it gives them a practical tool to evaluate, revise, and improve their classroom tests. Item analysis tells the teacher how students scored on each item. By using item analysis, a teacher can determine how well or how poorly each item in the test is functioning. The teacher can then revise the test by deciding which items to leave in the test and which items to delete or change. Two item statistics will be discussed later in this paper. They are item facility (IF) and the difference index (DI).

Consistency estimates

Consistency or dependability estimates for CRTs are comparable to the NRT notion of reliability. The central issue is the degree to which the teacher can expect the test to give the same results test after test. Because the main focus of this chapter is on item analysis, only one consistency estimate will be given here, the Kuder-Richardson formula number twenty-one (K-R-21). K-R-21 will be explained in more detail later.

Table 2. *Descriptive statistics*

Test	N	total possible	M	SD	Min	Max	Range
pretest	43	100	52.047	13.756	26	80	54
posttest	37	100	66.590	12.577	34	96	62

Results

The pretest results show that, of the fifty students enrolled, only forty-three took the pretest. The results also show a fairly wide spread of 54 points ranging from a low of 26 to a high of 80 points. The mean or average is about 52 points. Since the standard deviation was about 14 points and it is known that 34% of the test scores are one standard deviation plus or minus from the mean, 68% of the students scored from 38 to 66 points. Assuming the traditional pass-fail cutpoint at .70, 84% of the students failed on the pretest. This indicates that the test was effective in that the majority of students did not know the material when they entered the class at the beginning of the school year.

The posttest results show that thirty- seven students were still in the class when the final posttest was administered. The mean score increased from 52 to almost 67 points and the minimum and maximum scores also indicate some improvement. Another way to compare pretest and posttest scores is visually through the use of bar charts. You have already seen bar charts to show the normal distribution and skewed distributions. Figure 4 shows a bar chart in the horizontal view showing the distribution of pretest scores and Figure 5 shows a bar chart in the horizontal view showing the distribution of posttest scores. One student scored 70 on the pretest and 66 on the posttest. All other students improved from the pretest to the posttest.

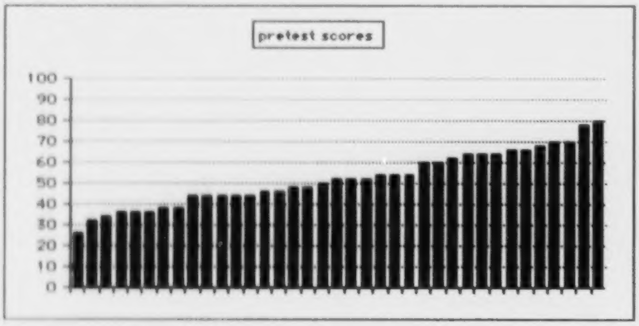


Figure 4. *Bar chart showing students' pretest scores.*

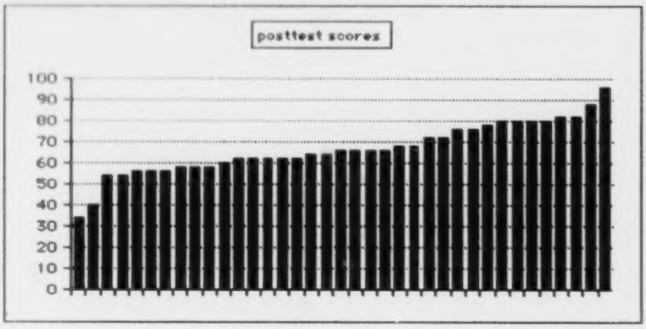


Figure 5. *Bar chart showing students and posttest scores.*

In this paper, only the item analysis statistics for items 1–15 are given.

Table 3. *Item statistics for items 1- 15.*

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Pretest															
IF	.70	.95	.59	.86	.70	.41	.76	.49	.73	.86	.03	.11	.79	.33	.49
Posttest															
IF	.86	.97	.70	.89	.57	.65	.86	.59	.89	.86	.03	.08	.94	.56	.70
DI	.16	.02	.11	.03	-.13	.24	.10	.10	.16	.00	.00	-.03	.15	.23	.21

Note: IF = item facility DI = Difference index

The item facility (IF) is an item statistic that gives the percent of correct answers (Brown, 1989). The formula is IF = N correct / N total. To determine the IF, divide the number of correct answers for an item by the total number of students who took the test. For example, on the pretest reported in this paper, the IF for item one was .70 and the IF for item eleven was .03. That means item one was answered correctly by 70% of the students and item eleven was answered correctly by only 3% of the students.

There are four possible outcomes for any test item in a pretest-posttest situation. These outcomes are given in Table 4.

As can be seen in Table 4, in outcome one, an item is answered correctly on the pretest and then answered incorrectly on the posttest. This is a rather bizarre situation because it means that the student knew the answer on the pretest and then forgot or in some way unlearned the answer for the posttest. In such a situation, the student may have been distracted while taking the test. Several things might distract students, and the teacher should investigate such occurrences. Distraction might come from an uncomfortable room, poor lighting, or taking the test the day after an all-night party. In outcome two of Table 4, an item is answered correctly on both the pretest and the posttest. In this case the item was too easy, probably because it was taught in previous courses. In outcome three, an item is answered incorrectly on both the pretest and the posttest. Perhaps this item was too difficult for the students to learn, or perhaps the teacher did not adequately review the item. It could also be the case that the teacher did not actually cover the item in class. In outcome four, the item was answered incorrectly on the pretest and answered correctly on the posttest. This is the ideal case for a CRT test item. The students came to the class not knowing this point, and due to their hard work (and good teaching) students exited the course knowing the item.

Table 4. *Possible CRT item outcomes*

possible outcome	pretest item	posttest item	explanation
1	correct answer	incorrect answer	student forgot or was distracted
2	correct answer	correct answer	item was too easy
3	incorrect answer	incorrect answer	item was too difficult, not taught, or not reviewed
4	incorrect answer	incorrect answer	ideal item for a CRT

To restate, the purpose of item analysis is to provide item statistics that enable the teacher to decide which of the four possible outcomes each item belongs to, so that the teacher can improve the test by deciding in future versions of the test which items to keep, which items to revise, or which items to reject. Let's see how this process works with the first fifteen items in our test. The IF index and how it was derived has already been explained. An IF index is calculated for each item in the pretest and each item in the posttest. The pretest IF is subtracted from the posttest IF and the resulting number is the Difference Index (DI). The formula is DI = posttest IF – pretest IF, and the higher the DI the better. The test was arranged in sections or groupings titled "tasks." The first fifteen items of the test include task 1 (items one through five), task 2 (items six through ten), and task 3 (items eleven through fifteen).

To begin the revision of task one, we see that the DI for items one through five are .16, .02, .11, .03, and -.13. These are not very impressive numbers. They indicate that item one showed an increase in scores of .16 or 16% which is not bad, but item two is only .02 or 2 percent and number five is a disaster with a minus sign indicating a net lose probably because some students experienced outcome one in Table 3. Looking at the actual test, items one through five appear as five blank lines. Students are instructed to listen and write the number they hear. The numbers the students hear on the tape are: four hundred, ninty-nine, eight thousand, thirty-two thousand five hundred, a hundred thousand and, a hundred and fifty thousand. Each number is repeated two times. The item analysis suggests that apart from item one, we could improve the questions, especially question five. Unfortunately, item analysis does not give us any idea of what to do. To improve the questions, we must use our knowledge and imagination. My goal was to make these items more difficult (so that more students would miss them on the pretest) and more easy (so that more students would get them correct on the posttest). What I decided, in fact, was to make the items easier by making the numbers smaller and to make the items more difficult by embedding them in a sentence. Item one was changed to, "Can you help me, I have to make twenty-five copies of this report." This sentence is taken directly from one of the units.

Items six through ten are also blank lines on the test sheet with instructions in print and on tape to "listen and write the prices you hear." Students hear various prices on the tape such as three dollars and seventy-five cents for item one and eighty-five cents for item ten. Item six seems to be functioning well with a DI of .24 but item ten has a DI of zero with a pretest IF of .86 and a posttest IF of .86 indicating that no learning took place probably because the item was too easy. In the revised test, all prices were embedded in sentences. For item six, students now hear, "Is fifty-four cents enough for an airmail letter to France?" and for item ten students hear, "A ticket for the bus is fourteen fifty and you can pay the driver."

Items 11 through 15 are multiple-choice questions. The students see four words, listen to the tape, and circles the word they heard. Items eleven and twelve were not functioning well (DI of zero and minus three) while items thirteen, fourteen, and fifteen are functioning much better. I interpreted the item analysis statistics as an indication to revise items eleven and twelve. Looking closely at item eleven, we see the four options: 1) you'll, 2) I, 3) I'll, and 4) this. The utterance students heard was, "You copy and I'll collate" which is taken from one of the dialogues in the textbook. Students are selecting answer one perhaps because it is easier to hear than the correct answer, number three. My revision was to keep the utterance and change distracter number one to a word not containing the sound /ou/ or /l/. By doing this, I reduced the distractor interference and students should find it easier to circle the correct answer. However, item analysis next year will confirm or deny my supposition. Next school semester, this test will again be administered as a pretest in April and posttest in January, and all items will be evaluated and revised as become necessary, especially items with a DI of less than .20.

Reliability is a statistical concept which is used to estimate inconsistency in a test. Imagine a perfect test. Such a test might exist in a Platonic heaven of perfect tests, but never here on earth. On earth all we have are imperfect tests all of which contain some inconsistency. We would generally like to know how reliable our imperfect test is.

This paper uses the Kuder-Richardson formula twenty-one or KR-21 which is an NRT statistic and, as such, is technically not appropriate for CRTs. However, there is an advantage to using KR-21. According to Brown (1990b), KR-21 is a conservative consistency estimate of the phi coefficient, a CRT statistic beyond the scope of this paper to explain. However, unlike the phi coefficient, KR-21 is easy to calculate because only three numbers are necessary, and they have already been calculated and reported in the descriptive statistics. They are the number of items, the mean, and standard deviation. The formula for KR-21 is

$$\text{K-R } 21 = \frac{k}{k-1}\left(1 - \frac{M(k-M)}{ks^2}\right)$$

where k = number of items
M = the mean of test items
s = standard deviation of the test items

In this calculation M and s are based on the sample posttest raw score data, and KR-21 turns out to be .85. k = 100, M = 66.59, and s = 13.35. This indicates that the students' scores are 85 percent reliable and 15 percent unreliable.

Discussion

In this paper, the distinction between NRTs and CRTs has been clearly illustrated. NRTs are of little or no help to classroom teachers in diagnosing their students' strong and weak points, assessing achievement, or evaluating programs. I have also shown how can CRTs can be designed and evaluated by using item analysis, which makes it possible to evaluate and improve the items. The results of the item analysis reported in this paper are sobering. The test described in this paper had been carefully designed (see Griffee, 1994, p. 34) by an English native speaker with an M.A. and many years of teaching experience. Despite these qualifications, many of the test items were shown by item analysis to be ineffective. This indicates that test items constructed by academically qualified and experienced teachers cannot be assumed to function as intended. Item analysis operates to flag certain test items and makes it possible for the teacher to identify those items in terms of one of the four outcomes in Table 4.

One weakness of this particular test is that the lack of institutional goals forced reliance on the textbook for test construction. Using a textbook instead of course objectives as the basis for the test raises two problems: one is the narrowness of the scope and the other is limitations of sequence. The problem of narrowness scope is that the focus of the test is restricted to a single textbook. In other words, lack of institutional program learning goals forces the teacher to make the textbook an end rather than a means toward an end. The problem of limited sequence is that, in any given course, the teacher has no way of relating or supporting other courses in the curriculum. The lack of institutional or departmental objectives means there is a risk that each course in the curriculum will become an isolated island with no bridges to the other islands.

Acknowledgements

An earlier form of this paper benefited from comments by K. Anderson and D. Reid. I am indebted to W. G. Kroehler for help in setting up the spreadsheet program.

References

Bachman, L. (1989). The development and use of criterioned-referenced tests of language ability in language program evaluations. In R. K. Johnson (Ed.), *The second language curriculum.*

Cambridge: Cambridge University Press (pp. 242-258).

Brindley, G. (1989). *Assessing achievement in the learner-centred curriculum*. Sydney: National Centre for English Language Teaching and Research.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23* (1), 65-83.

Brown, J.D. (1990a). Where do tests fit into language programs? *JALT Journal, 12* (1), 121-140.

Brown, J. D. (1990b). Short-cut estimates of criterion-referenced test consistency. *Language Testing* 7(1), 77-97.

Brown, J.D. (In press). *Testing in language programs*. Englewood Cliffs, N. J.: Prentice Hall.

Brown, J. D. (1992). Classroom-centered language testing. *TESOL Journal , 1* (4), 12-15.

Brown, J. D. (1993). A comprehensive criterion-referenced language testing project. In Douglas and Chapelle (Eds.), *A new decade of language testing research* pp 163-184. Washington D. C.: Teachers of English to Speakers of Other Languages.

Brown, J. D. & Pennington, M. C. (1991). Developing effective evaluation systems for language programs. In M.C. Pennington (Ed.), ***Building better English language programs. Perspectives on evaluation in ESL*** (pp. 3-18). Washington D C: NAFSA.

Griffee, D. T. (1992). *More HearSay: Interactive listening and speaking*. Reading, Mass. Addison-Wesley.

Griffee, D. T. (1994). Criterion-referenced test construction: A preliminary report. *Journal of Seigakuin University 6*, 31-40.

Richards, J., Platt, J., and Platt H. (1992). *Dictionary of Language Teaching & Applied Linguistics*. (2nd ed.). London: Longman.

Savignon, S. J. (1993). *Communicative competence: Theory and classroom practice*. Reading, MA. Addison-Wesley.

Chapter 4

Behavioral Learning Objectives as an Evaluation Tool

Judith A. Johnson
Kyushu Institute of Technology

If a poll were taken, probably most of the teachers around the world would say that they do not use the *Taxonomy of Educational Objectives* (Bloom, 1956) as an instructional aid. It is more than likely, however, that their teaching is in some way influenced by it—in the structure of the curriculum or syllabus they follow, the design of the textbooks they use, or the construction of the standardized tests they administer. The *Taxonomy* was originally conceived as an "educational-logical-psychological" system for classifying test items (p. 6) and "a method of improving the exchange of ideas and materials among test workers as well as other persons concerned with educational research and curriculum development" (p. 10). It was a very progressive schema in education at the time of its conception and quickly became a valuable tool for helping educators be more systematic in organizing educational and instructional objectives and assisting learners to develop "intellectual abilities and skills" (p. 38).

In addition to having an impact on large-scale achievement testing and curriculum evaluation, it has directly influenced classroom instruction and assessment as well. This influence was largely brought about by (a) the introduction of *instructional* and *behavioral objectives* and (b) the categorization of *lower-order* and *higher-order objectives*. In this chapter, first, the *Taxonomy* will be described; next, brief instructions on how to write behavioral objectives will be given; and then, explanations of how the use of objectives can improve planning, instruction, and evaluation of student learning will be presented. Finally, concerns about the use of the Taxonomy will be addressed, and its major contributions to classroom instruction and assessment will be discussed.

A Description of the Taxonomy

The *Taxonomy of Educational Objectives* (Bloom, 1956) is a list of six major classes of educational outcomes, based on the idea that the outcomes are hierarchically related as follows:

1.00 Knowledge
2.00 Comprehension
3.00 Application
4.00 Analysis
5.00 Synthesis
6.00 Evaluation

The most basic outcome being knowledge and the most complex being evaluation. In other words, the lower-order categories are prerequisite to achieving the higher ones.

Rohwer and Sloane (1994) identify six presuppositions about the nature of human learning and thinking upon which the structure and principles of the *Taxonomy* depend:

1. There is a relationship between learning and performance.
2. There are qualitatively different varieties of learning.
3. Learning is hierarchical and cumulative.
4. Learning is transferred to higher levels (vertically) and horizontally.
5. Generalized higher-order skills and abilities can be applied across content areas.
6. There is a difference in the way novices and experts learn new material.

Behavioral objectives came into existence when the authors of the *Taxonomy* decided that "virtually all educational objectives when stated in behavioral form have their counterparts in student behavior. These behaviors, then could be observed and described, and the descriptions could be classified" (Bloom, 1994, p. 3). The use of behavioral objectives enables both the teacher and the learner to see what is expected of the learner.

The cognitive structure of the *Taxonomy* was perceived to consist of lower-order and higher-order objectives. Although the division between lower-order and higher-order objectives has not been agreed upon by researchers, Bloom (1963) referred to Analysis, Synthesis, and Evaluation as classes which required "higher mental processes." Because of the differences between the two levels of objectives, different teaching methods for each have been proposed and supported by research findings. Teaching methods such as lectures have been found to be more effective in teaching lower-level objectives while communication-oriented methods such as discussions and group-work are more helpful in learning higher-order objectives (McKeachie, 1963; McKeachie & Kulik, 1975; Johnson & Johnson, 1991). Research into the two levels of objectives and teacher questioning practices has found a relationship between the types of questions asked by teachers and student achievement. When the greater number of questions posed by teachers are higher-order questions, student achievement is positively effected (Redfield & Rousseau,

Table 1. *Classes of Instructional Objectives and Corresponding Behaviors*

Knowledge— *behaviors*	the ability to remember previously learned information *define, name, state, reproduce, match, list, identify, describe*
Comprehension— *behaviors*	the ability to grasp the meaning of material *distinguish, defend, predict, explain, estimate, give examples, infer, paraphrase, summarize*
Application— *behaviors*	the ability to use learned material in new and concrete situations *use, modify, predict, compute, demonstrate, prepare, produce, show, discover, manipulate*
Analysis— *behaviors*	the ability to break down material from its component parts so that its organized structure can be understood *differentiate, relate, subdivide, separate, break down, identify, diagram, illustrate*
Synthesis— *behaviors*	the ability to put parts together to form a new whole *categorize, modify, tell, write, rewrite, summarize, generate, plan, reconstruct, explain, create, design*
Evaluation— *behaviors*	the ability to judge the value of material for a given purpose *compare, conclude, support, justify discriminate, critique, contrast, interpret, relate*

1981). Unfortunately, it appears that most teachers focus on lower-level questions (Anderson, Ryan, & Shapiro, 1989).

Table 1 is a summary of *Taxonomy* definitions of each class and corresponding observable behavior terms that can be used for stating specific learning outcomes.

Writing Behavioral Objectives

Behavioral objectives changed the focus of instruction from what the teacher will teach to what the student should learn. The objectives identify learner achievement in terms of behaviors which can be observed and measured by both the teacher and learners during and at the end of the instruction period. Comparing the following two objectives, the second objective is more clearly stated than the first.

1. Teach students how to write a business letter in English using an acceptable format
2. Students will write a business letter in English, using the format studied in class.

A minimum level of acceptable performance must be determined and announced to the learners. It should be included in the objective:

1. Write a business letter in English, *without errors*, using the format studied in class.

Any other important conditions that need to be known must also be stated:

1. *Using information provided*, write a business letter in English, without errors, *in the format studied in class. Dictionaries can be used.*

Well-stated objectives include the *learner* as the person who will produce a *visible action* that corresponds to the instructional objective. The level of *acceptable proficiency* and the *conditions* under which the learner must perform the action should also be included.

The first step in writing behavioral objectives is to determine the desired instructional outcome level. For example, if the objective is for learners to understand the content of a reading assignment, the instructional objective is at the Comprehension level: By the end of the course, the students will be able to comprehend the meaning of written material.

Next, behaviors (i.e., visible actions) which correspond to this level are identified and selected taking into consideration the nature of the reading materials, class size, length of instruction period, and other related factors.

By the end of the course, the students will be able to comprehend the meaning of the reading passage in Lesson 4 as demonstrated by their ability to:

1. Summarize the ideas
2. Explain the main ideas
3. Predict the outcome
4. Define key vocabulary items

Finally, the form of the behavior (written, oral, graphic, etc.) must be decided by the teacher, based on the curriculum objectives. (For more information on writing objectives see Mager, 1962; Popham, 1973; Gronlund, 1978; or, Brown, 1995).

Also, learning activities should be selected based on the potential of the activities to further curriculum objectives and benefit a specific group of learners. Lower-level behavioral objectives can be included but objectives from cognitive levels higher than the level being taught, generally, should not be used as they are probably too difficult for the learners. Mastery of lower-level objectives prepares learners to achieve higher-level objectives.

Planning Teaching/Learning Activities

Research related to lesson planning reveals that many teachers plan their lessons based on what they want the students to do rather than on what they want the students to learn (Peterson & Clark, 1986). In other words, they decide to use learning activities such as role-playing, writing a composition, or discussing a reading assignment without, first, identifying the specific aim(s) of the activity. As a result, before using an activity, teachers often fail to consider how the learners' acquisition of a new skill or knowledge will be evaluated. So at the end of the class, either no assessment or only very cursory (and perhaps subjective) assessment is made.

Using objectives in planning enables the teacher to see the inter-relatedness of all the phases of the instructional process (planning → teaching/learning → evaluation) from the outset. A commonly stated instructional aim for a foreign language class is, "Have students write an essay about hobbies." First of all, notice that the objective is for the teacher, not the learner. Secondly, note that the mere act of writing is considered the objective of the lesson. The purpose for writing the essay is unknown. In reality, writing the essay is not an objective, but an activity that provides the opportunity for students to learn, practice and apply, for example, specified language skills, composition rules and vocabulary. More accurately-stated *objectives* would be as follows:

(Application Level)

1. (The learner) uses the verb tenses studied in Lessons 3-6 with 90% accuracy when writing about given topics.
2. Uses specified punctuation with 80% accuracy when writing about given topics.
3. Uses given sentence structures related to hobbies with 90% accuracy.

In these objectives, it is understood that at the end of the instruction period, the learner is expected to use specific verb tenses, sentence structures and punctuation when writing about hobbies. With these objectives in mind, appropriate *activities* can be planned. For example:

1. (Learners) listen to a recording of speakers discussing hobbies and identify the hobbies mentioned
2. Make a list of hobbies (see who can make the longest list)
3. Say which hobbies they think are interesting
4. Write a paragraph explaining their hobby (or interest), or saying when they began it, why they like it, or how much time they spend on it when they do it.
5. Correct another student's paragraph
6. Ask others about their hobbies
7. Fill in punctuation in given essays

When clearly-stated objectives are used, lesson planning becomes an efficient and logical process, rather than a hit-or-miss affair. Equally important, the teacher and the learners can see the relationships among the behavioral objectives, the activities carried out in class, and the assessment of their achievement.

The concept of ***mastery learning***, developed by Bloom, is an instructional strategy founded on the belief that students can attain a specified level of learning if they are provided information and practice which is presented in a logical format and provides the learning time they require. According to the needs of the students, instruction can be individualized, carried out in small groups, or include the entire class. The use of the mastery learning strategies has repeatedly been proven to be highly successful. When the mastery approach is employed, an appreciably larger percentage of the learners have attained set objectives (Chung, 1994). And remember, the *Taxonomy* is the foundation of mastery learning instruction.

Learning and Assessment

Typically, learners cannot estimate the progress they've made in learning a foreign language until they receive their test results. This is because, generally, learners are not given observable standards by which they can measure their language abilities. Behavioral objectives focus on visible evidence that a skill has been learned at an acceptable standard, under given conditions. With this knowledge, learners can gauge their own progress at any given time. For example, if the objective is for the learner to use the past tense with a minimum of 80% accuracy when telling someone about a past experience, the learner can practice by recording accounts (or conversations with another person) of past experiences. The recording can be replayed and the verbs checked for accuracy. Learners can check their own work and/or each others' work. Needless to say, the learners should be given adequate time during the class to practice each skill. While the learners are practicing, the teacher has the opportunity to give them feedback on their progress. Similarly, objectives for reading, writing, and listening can be evaluated during the practice period. In this way, assessment becomes an integral part of the learning phase of the instructional process.

Criterion-referenced testing (CRT), used in the evaluation stage of the instructional process, is closely linked to the use of instructional objectives. A criterion-referenced test, as defined by the *Longman Dictionary of Language Teaching & Applied Linguistics* (Richards, Platt, & Platt, 1992, p. 91), is "a test which measures a student's performance according to a particular standard or criterion which has been agreed upon. The student must reach this level of performance to pass the test, and a student's score is therefore interpreted with reference to the criterion score, rather than to the scores of other students." Docking (1986) explains that "the base of criterion-referenced assessment is not the syllabus but objectives. The syllabus is derived from the objectives domain in the same manner as are the tests. In this way the tests do not just assess what was taught, rather they test whether or not the students have achieved the objectives the teaching (syllabus) was intended to facilitate." In norm-referenced testing (NRT), a student's score obtained on standardized tests is compared with the scores of all the other learners. In CRT, scores from tests (generally made by the teacher) provide information about how well the learner can perform specified skills and tasks or master a given content domain, and the scores are compared to an established standard. CRT can also be used to determine the effectiveness of specific objectives, materials, learning activities, and teaching methods. As Brown (1992) points out, the primary purpose of classroom assessment is "to foster learning." To this end, criterion-referenced test scores supply teachers with valuable information that can be used to improve the instructional process (Brown, 1990). Significant beneficial effects that CRT can have on educational programs are identified in Table 2.

Table 2. *Beneficial effects of criterion-referenced tests on students, teachers, and curriculum (adapted from Brown, 1992)*

Category	Beneficial Effects
Students	Diagnose strengths and weaknesses Motivate to study and review Reward for hard work
Teachers	Focus teaching efforts where needed Review areas of weakness Evaluate effectiveness of teaching
Curriculum	Reassess needs and objectives Revise and improve tests Modify materials and teaching

For the purpose of developing criterion-referenced evaluation tools such as achievement tests, quizzes, and random individual spot checks, behavioral objectives can be defined in terms of tasks. Sample criterion-referenced test items are given in Table 3. When selecting items as an assessment tool, the items should be based on the instructional objectives that were used to develop the syllabus and be categorized by cognitive levels so the teacher can be sure that the items cover what the learners are expected to know and yet are balanced with reference to difficulty.

CRT focuses on the learner's formative language development, permitting periodic self-assessment in addition to teacher assessment. Learner performance can be assessed in a variety of ways. Two common ways are as follows: (a) indicating whether or not the student meets the minimum level of accuracy required (e.g., can/cannot use the past tense), (b) distinguishing levels of mastery (e.g., uses past tense with infrequent/frequent errors, but meaning is/isn't conveyed clearly using past tense). As many teachers use a point system for assessing learners' performance, points can be assigned to the levels of mastery of each objective.

Table 3. *Sample criterion-referenced test items*

(Knowledge)
Know the meaning of vocabulary items.
1. Select the picture that expresses the meaning of ______.
2. State (say or write) a synonym for _______.
3. Write a definition for each of the words below.

(Comprehension)
Understand written material.
1. Write a summary of the paragraph, below.
2. Give (say or write) two more examples of the type of _______ discussed in the passage.
3. Explain (say or write) why the main character of the story ________.

(Application)
Follow oral directions.
1. Draw a line on the map along the route you are told to take by the speaker.
2. After listening to a series of directions you are given, identify your destination.

Give oral directions.
1. Guide someone to a place he/she wants to go in this school.
2. Give someone directions on how to get from __________ to the place he/she wants to go to near ____________.

(Analysis)
Identify the organizational structure of a composition.
1. Write an outline of the organization of the main and supporting ideas of the composition, below.

Point out unstated assumptions in given literary works.
1. After reading the following article, give (say/write) three things that the author assumes the reader accepts as truths.

(Synthesis)
Combine information from different sources to tell about a topic.
1. In a five-minute speech, tell the class about _______, using information obtained from at least three different sources.
2. Write a play, story, or narration about ________, using parts of personal accounts you've heard from at least two people who've experienced _________.

(Evaluation)
Critique a work of art.
1. Give (say or write) your opinion about the film, ________ citing specific incidents and facts to support your ideas.
2. Compare the ideas presented in ___________ with what we know, today, as scientific fact related to this topic.

Discussion

While most educators, to varying degrees, accept the assumptions upon which the *Taxonomy* is based (see Gagné, 1977; Rohwer & Sloane, 1994), some contend that learning is not necessarily hierarchical. Others (Moore & Kennedy, 1971; Purves, 1971; Orlandi, 1971) point out that in some subject areas, comprehension is enhanced when the categories are ordered differently. Madaus, Woods, and Nuttall (1993) found the levels of synthesis and evaluation to be similar. However, other investigations by Ekstrand (1982) and Hill

(1984) yield data that favor the cumulative-hierarchy theory. At this time, however, findings are still inconclusive.

Some teachers (Eisner, 1967; Brokehoff, 1979; Tumposky, 1984) have taken a very narrow view of the use of behavioral objectives in language teaching. Tumposky (1984) proposes that, due to the unpredictability and creativity involved in foreign language learning, objectives are not suitable for this type of instruction. However, the structural, phonetic, and social aspects of language learning are replete with predictable language patterns at all speaking levels. In fact, using higher-order objectives may also help students to exercise their creativity and their ability to express more sophisticated ideas. Objectives can bring clarity and purpose to activities and tasks students are often asked to perform in intermediate and advanced foreign language classes. Rather than merely receiving vague instructions to discuss a topic or write a composition on a given theme, the use of objectives enables learners to know the specific aims of the assignments.

Another issue raised by opponents of behavioral objectives is that not all types of learning have observable outcomes. This may be true. However, in foreign language learning, where using the target language is of primary importance, almost every learning outcome will have a corresponding behavior that can be described. *Handbook II*, the second volume of the *Taxonomy* (Krathwohl, Bloom, & Masia, 1964), identifies objectives in the affective domain, a domain which also greatly influences foreign language acquisition.

Some critics are afraid that the use of objectives will result in taking the spontaneity out of language learning (and teaching), especially if teachers become preoccupied with insignificant aspects of the language, dividing it into many isolated parts that are difficult to relate to actual language use. Foreign language learners in Japan would undoubtedly be of the opinion that this is the current state of most foreign language instruction. What they are learning in most language classes is remote from the ways in which language is used in normal, daily life. Viewed from the functional and notional aspects of language acquisition, behavioral objectives can help teachers plan instruction that would result in practical, meaningful communication-oriented learning outcomes (van Ek & Alexander, 1977; Johnson, 1994).

Those who are unfamiliar with the *Taxonomy* have said it's too complicated to use and, therefore, requires a lot of planning time. Most skills take time to be learned. Foreign language students are expected and required to devote quite a lot of out-of-class time to the subject. Should teachers shortchange their students because they don't want to invest extra time in learning how to provide better instruction to the students? Actually, once learned, the very logical and systematic nature of the *Taxonomy* makes all phases of the instruction cycle easier to execute.

The difficulty in using the *Taxonomy* to classify objectives (which the authors also acknowledged) is that the teacher must know or rely on assumptions about the learners' prior educational experiences. Most teachers would probably agree, though, that this difficulty is not unique to the *Taxonomy*. CRT, which is most suitable for evaluating behavioral outcomes according to Bachman (1989), is not yet able to define language proficiency levels because the content of a criterion-referenced test must be taken from a "well-defined domain of ability," in which there must be "an absolute scale of ability" on which learners' performance can be measured (pp. 255-256). The upper and lower limits of foreign language proficiency and the standard against which the foreign language learner can be measured have yet to be determined. Although research is being carried out in the development of language competency scales and models, the difficulties being faced are the same as those found in present instruments used to evaluate language proficiency: terminology is relative, distinct levels of proficiency cannot be determined, ratings are biased by oral evaluators, and the like.

Some evidence exists that the ability to understand the events which take place in the classroom and the skill to guide classroom activities are important elements of a teacher's success in classroom management (Doyle, 1986). If teachers can identify the behavior they expect the learners to exhibit at the end of a lesson, planning the lesson becomes a logical process (Gower & Walters, 1983). Using objectives to design the course syllabus, select activities, and create evaluation instruments yields multiple sound results: the program of instruction can be logical and cohesive; teachers and students can clearly know what is to be learned and how that learning is to be assessed; once learners understand what they must do, they can take responsibility for their own learning and self-assessment; assessment of the learner's progress can take place during the learning process; a variety of teaching methods and techniques can be used to help students achieve objectives; teachers and students can focus on using the language, skill, material, etc. being studied; teachers can determine the general cognitive level of their instruction (which can be changed, if necessary); and teachers' classroom management skills can be enhanced (see Brown, 1995, for more on this topic). It is evident that the use of objectives can reap numerous benefits for teachers. However, it is important to remember that the central reason for using behavioral objectives is to improve students' learning.

Educators involved in language teaching are obligated to provide foreign language learners the best possible learning environment, instruction, and assessment of their language skills. The intelligent use of learning objectives combined with other sound educational practices appears to be the optimum way to fulfill this responsibility.

References

Anderson, D., Ryan, W., & Shapiro B. J. (Eds.). (1989). *The IEA classroom environment study*. Oxford: Pergamon.

Bachman, L. F. (1989). The development and use of criterion-referenced tests of language proficiency in language program evaluation. In R. K. Johnson (Ed.), *The second language curriculum*. Cambridge: Cambridge University Press.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives, handbook I: Cognitive domain*. New York: David Mckay.

Bloom, B. S. (1963). Testing cognitive ability and achievement. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally.

Bloom, B. S. (1994). Reflections on the development and use of the taxonomy. In L. W. Anderson & L. A. Sosniak (Eds.), *Bloom's taxonomy: A forty-year retrospective*. Chicago: University of Chicago.

Broekhoff, M. (1979). Behavioral objectives and the English profession. *English Journal, 68*(6), 55-59.

Brown, J. D. (1990). Where do tests fit into language programs? *JALT Journal, 12*(1), 121-140.

Brown, J. D. (1992). Classroom-centered language testing. *TESOL Journal, 1*(4), 12-15.

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development*. Boston: Heinle & Heinle.

Chung, B. M. (1994). The taxonomy in the Republic of Korea. In L. W. Anderson, & L. A. Sosniak (Eds.) *Bloom's taxonomy: A forty-year retrospective*. Chicago: University of Chicago.

Docking, R. A. (1986). Norm-referenced and criterion-referenced measurement: A descriptive comparison. *Unicorn, 12*(1).

Doyle, E. (1986). Classroom organization and management. In M. C. Wittrock (Ed.) *Handbook of research on teaching*. New York: Macmillian.

Eisner, E. W. (1967). Educational objectives: help or hindrance? *School Review, 75*(2), 250-260.

Ekstrand, J. M. (1982). Methods of validating learning hierarchies with applications for mathematics learning. Paper presented at the Annual Meeting of the American Educational Research Association, New York City, 1982. (ERIC Document Reproduction Service ED 216 896).

Gagné, R. M. (1977). *The conditions of learning*. New York: Holt, Rinehart, and Winston.

Gower, R., & Walters, S. (1983). *Teaching practice handbook*. Oxford: Heinemann.

Gronlund, N. E. (1978). *Stating objectives for classroom instruction*. New York: Macmillan.

Hill, P. W. (1984). Testing hierarchy in educational taxonomies: A theoretical and empirical investigation. *Evaluation in Education, 8*, 179-278.

Johnson, E. W., & Johnson, R. T. (1991). Classroom instruction and cooperative learning. In H. C. Waxman, & H. J. Walberg (Eds.), *Effective learning: Current research*. Berkeley, CA: McCutchan.

Johnson, J. A. (1994). Using behavioral learning objectives to improve the teaching of scientific subjects. *The Bulletin of the Faculty of Computer Science and Systems Engineering Kyushu Institute of Technology, 7*, 99-107.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives, handbook II: Affective domain.* New York: Longman.

Madaus, G., Woods, E. M., & Nuttall, R. L. (1993). A causal model analysis of Bloom's taxonomy. *American Educational Research Journal, 10*, 253-262.

Mager, R. F. (1962). *Preparing instructional objectives.* Palo Alto, CA: Fearon.

McKeachie, W. J. (1963). Research on teaching at the college and university level. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally.

McKeachie, W. J., & Kulik, J. A. (1975). Effective college teaching. In F. N. Kerlinger (Ed.), *Review of research in education.* Itasca, IL: Peacock.

Moore, W. J., & Kennedy, L. D. (1971). Evaluation of learning in the language arts. In B. S. Bloom, J. T. Hastings, & F. Madaus (Eds.), *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill.

Orlandi, L. R. (1971). Evaluation of learning in secondary school social studies. In B. S. Bloom, J. T. Hastings, & F. Madaus (Eds.), *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill.

Peterson, P. L., & Clark, C. (1986). Teachers' thought processes. In J. C. Wittrock (Ed.), *Handbook of research on teaching.* New York: Macmillan.

Popham, W. J. (1973). *The uses of instructional objectives.* Belmont, CA: Fearon.

Purves, A. C. (1971). Evaluation of learning in literature. In B. S. Bloom, J. T. Hastings, & F. Madaus (Eds.), *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill.

Redfield, D. & Rousseau, E. W. (1981). A meta-analysis of experimental research on teacher questioning behavior. *Review of Educational Research, 51*, 244.

Richards, J. C., Platt, J., & Platt, C. (1992). *Longman dictionary of language teaching & applied linguistics.* Essex: Longman House.

Rohwer, W. D., & Sloane, K. (1994). Psychological perspectives. In L. W. Anderson, & L. A. Sosniak (Eds.) *Bloom's taxonomy: A forty-year retrospective.* Chicago: University of Chicago.

Tumposky, N. R. (1984). Behavioral objectives,the cult of efficiency and foreign language learning: Are they compatible? *TESOL Quarterly, 18*(2), 295-310.

van Ek, J. A. & Alexander, L. G. (1977). *Systems development in adult language learning: Waystage.* Strasbourg: Council of Europe.

Section II

Program-Level Testing Strategies

Chapter 5

Developing Norm-Referenced Language Tests for Program-Level Decision Making

JAMES DEAN BROWN
UNIVERSITY OF HAWAII AT MANOA

When I was an ESL and EFL teacher and administrator, I encountered two basic types of language tests: tests that administrators need in order to classify or group students in some way or other, and tests that teachers need to help them determine what their students have learned in a particular course. These two types of tests are usually called *norm-referenced tests*, and *criterion-referenced tests* (for more on this distinction, see Chapter 2 of this book, or Hudson & Lynch, 1984; Brown, 1984, 1988, 1989, 1990, 1992, 1993, 1995a, & 1995b; Bachman, 1990).

At the same time, I have long recognized that language tests are most commonly used to make six basic types of decisions in language programs: language aptitude, proficiency, placement, diagnostic, progress, and achievement tests. Because their basic purpose is to classify or group students, norm-referenced tests are best suited for making three of these types of decisions: aptitude, proficiency, and placement decisions.

For instance, language aptitude tests, like the one I took at the beginning of my military service, are used to decide who will most benefit from language training (or from the U.S. Army's point of view: who will be their best investment?). In contrast, general language proficiency tests are most commonly used for admissions tests (in the way that TOEFL scores are used to decide which international students should be admitted to American universities). Later, when students have already been admitted to a particular institution, it may be necessary to use placement tests to decide which levels of language study students should pursue (like the *English Language Institute Placement Test*, or ELIPT, used at the University of Hawaii at Manoa to decide whether students should study at the intermediate or advanced levels or be exempted from ESL study). In all three cases—aptitude, proficiency, and placement tests—the scores are needed to make decisions that compare each student to all other students so they can be classified (i.e., as a good or bad investment in the case of aptitude; as admitted or not admitted in the case of proficiency; or placed in levels of language study in the case of placement testing).

Most of the language testing literature that has developed over the years has been about norm-referenced language testing in one way or another (see Bachman, 1990; Brown, 1995a for overviews of the norm-referenced language testing literature). Yet, few papers have looked at norm-referenced testing as a set of tools for making decisions in language programs. In this chapter, I will provide answers to three basic questions about norm-referenced testing:

1. What are norm-referenced tests, and what are they used for?
2. What do administrators need in their norm-referenced tests?
3. How should norm-referenced tests be developed and improved?

Examining these three questions, one at a time, should help administrators and teachers to understand how important norm-referenced testing is, as well as why and how they should be doing their norm-referenced testing.

What Are Norm-referenced Tests, and What Are They Used for?

Norm-referenced tests will be defined here as those tests which are used in language programs to assess students' aptitude to learn a language, measure their proficiency, or determine their appropriate level of study in a particular language program.

In the language programs that I have known, the norm-referenced tests were either (a) adopted from other sources, (b) developed within the program, or (c) adapted to meet the needs of the program. As director of the ELI at the University of Hawaii, I was an administrator who sometimes chose to adopt already existing norm-referenced tests. For instance, we adopted a test for our overall proficiency testing called the *Test of English as a Foreign Language* (ETS, 1990). We also choose to develop our own tests for placement purposes. Our placement test was the *English Language Institute Placement Test* (see Brown, 1989a, 1995a, or 1995b for fuller descriptions). Since we had to deal with all students who were accepted by the university, language learning aptitude was not an issue for us so we have never done any aptitude testing.

At other institutions, administrators may choose to adopt placement tests, or adapt tests that are already in place in their institutions. The point is that putting sound norm-referenced tests into place is an important responsibility. This responsibility will typically fall on the administrator(s). Classroom teachers may be involved in norm-referenced testing as a sort of labor pool for administering and scoring the tests, but seldom are they directly involved in the selection, development, or adaptation of such tests. I am not saying that teachers should not take an interest in the norm-referenced types of decisions that are closely linked to program level decisions. I am just saying that teachers are likely to be most directly responsible for and most interested in the criterion-referenced types of decisions discussed in Chapters 2, 3, and 4 of this book.

Regardless of the level of interest they take in the norm-referenced tests, teachers must recognize that they benefit in a number of ways from norm-referenced tests. For example, because of aptitude, proficiency, and/or placement testing, teachers may have relatively homogeneous groups of students to teach in each of their classes. Such homogeneity may take the form of having only students with a relatively high degree of language learning aptitude in the class, or only students above a certain admissions level as determined by a proficiency test, or only students within a certain band of abilities as determined by a placement test, or all three. Once provided such homogeneity, teachers can carefully tailor their classroom activities, exercises, homework, and so forth to the needs of a clearly defined group of students. Any teacher who has ever had to teach students with a wide range of aptitudes or abilities will easily understand the value of this benefit—a benefit that comes from using sound norm-referenced testing practices.

What Administrators Need in a Norm-referenced Test

The question of what administrators need in a norm-referenced test is really a two-part question. First, what types of test information do administrators need? Second, what qualities are desirable in a norm-referenced test?

Types of test information. Naturally, there are many types of information that a norm-referenced test might usefully provide, and the types will vary from institution to institution and administrator to administrator. However, I most often find myself needing to know: whether the students' language proficiency is high enough to enter our institution, and what level (if any) of ESL study they should be placed into. These are questions that can't be answered on the basis of the information provided on typical criterion-referenced tests because such issues are very global and can only be addressed by classifying students into groups for comparison using well-developed norm-referenced tests.

Ideal norm-referenced test qualities. An ideal norm-referenced test is one that fits the group for which it was designed in terms of the difficulty of the items, and one that has items on it that discriminate well between the high ability students and the low ability students. In addition, an ideal norm-referenced test is one that efficiently spreads the students out along a continuum of abilities so that grouping decisions can be made easily and responsibly. Whether adopting, developing, or adapting norm-referenced tests, administrators still have to examine the tests very carefully and critically, just as they do with textbooks, to determine whether or not the tests are up to their standards.

Unfortunately, most programs are unable to find off-the-shelf tests that meet their needs, especially for purposes of placement. A given placement test is designed for making placement decisions within a particular population of students, and such a test may be totally inappropriate for the different population of students found in another program. As a result, programs often end up writing their own tests.

The norm-referenced test review checklist shown in Table 1 provides a list of crucial questions that should be addressed when developing such a norm-referenced test. If you go down the list and carefully address each of the questions, you should be able to produce a test of respectable quality.

Table 1. *Norm-referenced test review checklist*

[ ] Are the test questions that I have written directly related to the teaching materials and activities being used in class?

[ ] Are there enough test questions so that I can eliminate some bad ones.

[ ] Has at least one colleague critically looked over the test?

[ ] Is the test easily reproducible on locally available equipment?

[ ] If I have previously used the test, did I revise it on the basis of what happened in that administration?

[ ] If I have previously used the test, does that mean that it is available to some of the students?

[ ] Do I want to develop the test in multiple forms?

[ ] Are the directions concise and adequate?

[ ] Does the test have clear, complete, and correct answer keys and directions for scoring?

[ ] Is score interpretation clear?

[ ] Have provisions been made for clearly reporting the scores to the students?

[ ] Is the test of demonstrably good quality?

How Should Norm-referenced Tests Be Developed and Improved?

All tests should be developed such that the test items make sense in terms of the curriculum for which they are being designed. In other words, the items should be carefully written to reflect the teaching points and types of activities that go on in the classroom. Because these aspects of language teaching vary considerably, the decisions about the types of items to include will have to be made locally. However, there are several general approaches that can be used in developing and improving norm-referenced tests. Here, I will describe what I call the minimum process of norm-referenced test development, as well as what I call the full process of norm-referenced test development.

The minimum process of norm-referenced test development. In fact, the checklist shown in Table 1 can be used at various stages of the test development process to maximize the quality of your norm-referenced tests, and thereby minimize the probability of making errors in the aptitude, proficiency (particularly admissions), or placement decisions that you must make based on the tests. At minimum, you should include the following steps in your test development process:

1. Once the test is actually adopted, adapted, or developed, you should critique it using the checklist in Table 1.
2. Take the test yourself before you administer it. This will help you to spot problems and insure that you have an accurate answer key to work from.
3. Have a number of colleagues look over the test before administering it.
4. Take notes during the test administration on anything that the students ask about, or on any problems that you notice while correcting the tests.
5. Most important, revise the test immediately after administering it (while the problems are still fresh in your mind) and use all that you have learned in steps one to four above in the revision process.

If you follow these minimum steps and use the checklist provided in Table 1, you could find yourself in the enviable position of providing relatively sound norm-referenced tests for your students.

The full process of norm-referenced test development. To find out how well the test questions are functioning in a norm-referenced test, administrators must examine each question to find out how students perform on it. To do this effectively, the administrator must examine at least one set of test results gathered either by pilot testing it with a group of students like the ones who will eventually be tested, or by administering the test operationally and then analyzing the results.

On a question-by-question basis, several statistics exist that can help teachers to decide which test questions fit the ability levels of their group of students and discriminate well between the high and low ability students. The statistics explained here require only that a test be administered on one occasion.

Calculating the facility index. The *facility index* is one very easy statistic that can help teachers determine how well a particular test question fits the ability levels of the students involved. As mentioned in Chapters 2 and 3, the facility index is calculated by adding up the number of students who correctly answered a particular test question and dividing by the number of students who took the test; these calculations yield the proportion of students who answered correctly. For instance, if 40 (out of 200) students correctly answered a test question, the facility index would be 40/200 = .20. Understanding this index is easy because moving the decimal point two places to the right turns this index into a percent. In the example, .20 would become 20 percent. This specific question might be considered fairly difficult for the students in this group because only 20 percent answered correctly.

Interpreting the facility index. So interpreting facility indexes is relatively easy. For instance, a facility index of .95 indicates that 95 percent of the students answered the question correctly, and that it was a very easy question for the particular group of students who were tested. An item facility of .10 indicates that the question was very difficult for the students because only 10 percent could answer it correctly.

An ideal question for a norm-referenced test is one that 50-60 percent of the students answer correctly. Typically, testing books say that items can be retained in the norm-referenced test revision process if their facility values fall between .30 and .70 because items outside of that range really ought to be considered either too hard or too easy for the group being tested.

Calculating the discrimination index. Another straightforward statistic, called the *discrimination index*, is used to examine the degree to which the high ability, or top scoring, students on a norm-referenced test item answer it correctly as compared to the low ability, or bottom scoring, students. To calculate a discrimination index, you must first decide which students belong in the top scoring group and which belong in the bottom scoring group. Typically, the scores on the whole test are lined up from high to low and then the top and bottom third are chosen to be the top and bottom groups, as shown in Table 2. (Depending on the number of students, it may be more convenient to use the top and bottom 25 percent or something between 25 and 33 percent.) Once the top and bottom groups are decided, the discrimination index is calculated in three steps:

1. Calculate the item facility for the top group (see the row labeled IF_{TOP} in Table 2)
2. Calculate the item facility for the bottom group (see the row labeled IF_{BOT} in Table 2)
3. Calculate the item discrimination index (ID) by subtracting the item facility for the bottom group from the item facility for the top group ($IF_{TOP} - IF_{BOT} = ID$)

For example, item 1 in Table 2 has an item facility for the top group of 1.00 and an item facility for the bottom group of 1.00. In other words, everybody in the top and bottom groups answered this item correctly. Calculating the item discrimination index using the formula ($IF_{TOP} - IF_{BOT} = ID$), ID would turn out to be 1.00 - 1.00 = 0.00. In other words, this item has zero discrimination, which makes sense because the top and bottom groups did equally well on this item.

Let's consider several other examples. Item 2 in Table 2 has an item facility for the top group of 0.00 and an item facility for the bottom group of 0.00. In other words, everybody in the top and bottom groups answered this item wrong. Calculating the item discrimination index using the formula ($IF_{TOP} - IF_{BOT} = ID$), ID would turn out to be 0.00 - 0.00 = 0.00. In other words, this item has zero discrimination, which makes sense because the top and bottom groups did equally poorly on the item.

In contrast, on item 3 in Table 2, the item facility for the top group was 1.00 and that for the bottom group was 0.00. In other words, everybody in the top group answered this item correctly and everyone in the bottom group answered it incorrectly. Calculating

Table 2. *Calculating NRT Discrimination Indexes*

Student	Items										
	1	2	3	4	5	6	7	8	9	10	......Total
YOKO	1	0	1	0	1	1	1	1	1	1	 47
TOSHI	1	0	1	0	0	0	1	0	1	1	 44
SAYOKO	1	0	1	0	1	1	1	0	1	0	 42
HIDE	1	0	1	0	0	1	0	0	1	0	 41
NAOYO	1	0	1	0	1	1	1	1	1	1	 39
RIEKO	1	0	1	0	0	1	1	0	1	1	 35
KEIKO	1	0	0	1	1	1	1	0	0	1	 34
NAOMI	1	0	0	1	0	0	1	0	0	1	 30
TAEKO	1	0	0	1	1	1	1	0	1	0	 30
YUKIE	1	0	0	1	0	0	0	0	1	0	 30
ASAKO	1	0	0	1	1	0	0	0	0	0	 27
HIROTO	1	0	0	1	0	0	0	0	1	0	 13
IF	1.00	0.00	0.50	0.50	0.50	0.58	0.67	0.17	0.75	0.50	
IF_{TOP}	1.00	0.00	1.00	0.00	0.50	0.75	0.75	0.25	1.00	0.50	
IF_{BOT}	1.00	0.00	0.00	1.00	0.50	0.25	0.25	0.00	0.75	0.00	
ID	0.00	0.00	1.00	-1.00	0.00	0.50	0.50	0.25	0.25	0.50	
			#			#	#	•	•	#	

the item discrimination index using the formula (IF_{TOP} - IF_{BOT} = ID), ID would turn out to be 1.00 - 0.00 = 1.00. This value of 1.00 is as high as the discrimination index can go. In other words, this item has perfect discrimination because it discriminated as well as is possible between the top and bottom groups (i.e., everyone in the top group did as well as possible and everyone in the bottom group did as poorly as possible).

Item 4 in Table 2 shows what happens when for some reason the students in the bottom group all get the item right (1.00) and everybody in the top group answers it incorrectly (0.00). In this case, ID would turn out to be 0.00 - 1.00 = -1.00. This value of -1.00 is as low as the discrimination index can go. It indicates that the item is doing something completely opposite from the test as a whole (i.e., everyone in the top group did as poorly as possible and everyone in the bottom group did as well as possible).

However, these first four examples are extreme examples because they were designed to illustrate the values that a discrimination index can take on. If a set of norm-referenced test items are well written, they will typically take on intermediary values more like those shown in items 5 to 10 in Table 2.

Interpreting the discrimination index. Remember, the point of item statistics is to select those items that discriminate best for a revised version of the test. Deciding which items to keep and which to eliminate will depend in part on how many items you need to keep. Let's say that four items are needed from this first batch of 10. The decision would then involve examining the ID values and keeping those items that, in this case, have IDs of .50 or higher (items 3, 6, 7, & 10, that is, the ones with a # at the very bottom). If six items were needed then the standard for ID might drop considerably to .25 to include items 8 and 9 (the ones with an * at the very bottom).

You should also keep an eye on the average IF for the items that remain on a revised version of the test, because it indicates how difficult the test has become in the revision process. For instance, in the six item revision in the above example, the average IF would be (.50 +.58 +.67 +.17 +.75 +.50)/6 = .53. In other words, the average percent answering the items correctly was 53 percent. Of course, if item 8 (with its IF of .17) is also eliminated from the test, the resulting average IF for the remaining five items will be (.50 +.58 +.67 +.75 +.50)/5 = .60. Since this indicates that 60 percent on average are answering correctly, the test will be considerably easier than the six item version with its average IF of .53. You will simply have to decide how easy the test should ultimately be.

In short, selection of items in any norm-referenced test revision process must also take into account common sense and practical considerations like how long the test must ultimately be and how difficult it should be. As a result, the only rule of thumb I can give you for interpreting ID in a norm-referenced testing situation is to try to keep those items which have the highest item facility while keeping a sharp eye on the quality of the items themselves, the number of items that are needed, and the difficulty of the resulting test.

To expand the minimum process of norm-referenced test development given earlier, a number of steps can now be added, especially the steps that examine how students are performing on the individual test questions. The full process would thus include at least the following steps:

1. To improve test validity, you should critique the test using the checklist in Table 1 with special attention to checking that the test questions match the sorts of things taught in your program.
2. Take the test yourself before you administer it. This will help you to spot problems and insure that you have an accurate answer key to work from.
3. Have a colleague look it over before administering it.
4. Take notes during the test administration on anything that the students ask about, or on any problems that you notice while correcting the tests.
5. Compile the results (from a piloting of your test, or from an operational adminis-

tration) on an question-by-question basis.

6. Code each student's answers in such a way that you can calculate the percent of students who answered each test question correctly.
7. Calculate overall facility indexes (as shown in Table 2).
8. Arrange your item data so that they are tabulated according to the students' total scores, from high to low.
9. Identify a top and bottom scoring group consisting of about 33 percent of the students each.
10. Calculate the item facility index for each group (as shown in Table 2).
11. Calculate a discrimination index for each item by subtracting the item facility for the bottom group from the item facility for the top group.
12. Interpret the facility and discrimination indexes in terms of their ramifications for revising the particular test involved while keeping in mind the quality of the actual items, the number of items that are needed, and the difficulty of the resulting test.
13. Most important, revise the test immediately after administering and analyzing it (while the problems are still fresh in your mind) and use all that you have learned in steps one to twelve above.

If you follow these 13 steps, you will have created (a) a better quality test, (b) a test that is more closely related to your students' ability levels, and (c) a test that spreads your students out more efficiently so that you can make more responsible norm-referenced decisions such as the admissions and placement decisions we make regularly at the University of Hawaii at Manoa.

Conclusions

This chapter began by defining norm-referenced tests and listing some of the different types of information that can be gathered using norm-referenced tests. It also provided a checklist for reviewing the quality of NRTs and explained the processes involved in developing and revising them—including two statistics, the facility and discrimination. In short, quite a bit of ground was covered in this discussion of NRTs, yet the subject is far from being exhausted.

The single greatest problem in norm-referenced language testing is that administrators don't recognize that most off-the-shelf tests may not be suitable for their particular language program. NRTs are developed using item facility and discrimination statistics to fit a particular group of students. Since groups of students vary widely from one institution to another, it may be totally inappropriate to use a norm-referenced placement test developed and published by one institution at another institution. For instance, as director of the ELI at University of Hawaii at Manoa, I often got calls from directors of other language programs in Hawaii (and elsewhere) asking if they could copy or buy our placement test. My answer was always a resounding, though polite, "NO!" for two reasons: (a) I did not want to compromise the security of our six-test battery, which involved a great investment of time, knowledge, energy, and money; and (b) our tests were tailored by item analysis to fit our group of students (who ranged from about 500 to 600 on the TOEFL), and thus our tests would be totally inappropriate at most other institutions, where the ranges of ability are typically quite different.

Unfortunately, many administrators do not have the time or knowledge to build in-house NRTs that will be suitable for their programs. To address this issue, strategies must be developed for finding or training people who will have the time and knowledge to develop in-house NRTs that will help in making aptitude, proficiency (particularly admissions), and placement decisions as necessary. If you are too busy to apply the information supplied in this article, there are three other strategies you might try: (a) hire a teacher who also is trained in the important area of language testing and release that person from some teaching duties so that he/she can develop effective in-house tests; (b) identify one teacher, who is interested in testing, and send that teacher to one of the many ESL/EFL teacher training programs around the world where language testing courses are offered (for instance, the TESOL Summer Institutes—one or two of these occur every year); or (c) hire a testing specialist to come briefly to your program to work with and train selected staff members (perhaps in a hands-on manner by helping them to develop tests for your program).

In short, NRTs must be accorded a much higher priority in the administrator's program development plans. Important decisions about students' lives are made on the basis of these tests. Hence, they must be accorded a more important place in language curricula everywhere. To do so, language programs must marshal increasing resources in terms of time, money, materials, and training so that administrators and teachers can develop and use norm-referenced testing as an important, effective, and integral part of their decision making processes.

In essence, resources must be found to help people like you work on your NRTs. Given even minimal support, you can develop sound NRTs that can in turn help you to make more effective decisions. Norm-referenced tests like the university entrance examinations in Japan, which may or may not be of good quality (see Brown & Yamashita in Chapter 10), are already a reality to most students in Japan, and that is all-the-more reason why you must take responsibility for creating sound norm-referenced tests in your language program so that at least the aptitude, admissions, and placement decisions that you make within your program will be efficient, effective, and fair to your students.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design.* Cambridge: Cambridge University.

Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly, 23*(1), 65-83.

Brown, J. D. (1990). Where do tests fit into language programs? *JALT Journal, 12*(1), 121-140.

Brown, J. D. (1992). Classroom-centered language testing. *TESOL Journal, 1*(4), 12-15.

Brown, J. D. (1993). A comprehensive criterion-referenced language testing project. In D. Douglas and C. Chapelle (Eds.) *A new decade of language testing research* (pp. 163-184). Washington, DC: TESOL.

Brown, J. D. (1995a). *Testing in language programs.* Englewood Cliffs, NJ: Prentice-Hall.

Brown, J. D. (1995b). *The elements of language curriculum: A systematic approach to program development.* New York: Heinle & Heinle.

ETS. (1990). *Test of English as a foreign language.* Princeton, NJ: Educational Testing Service.

Hudson, T., & Lynch, B. (1984). *A criterion-referenced approach to ESL achievement testing. Language Testing, 1*(2), 171-201.

Chapter 6

Monitoring Student Placement: a Test–Retest Comparison

SAYOKO OKADA YAMASHITA
INTERNATIONAL CHRISTIAN UNIVERSITY

Many Japanese language education institutions either use the *Test of Japanese Language Proficiency*[1] or a placement test that they developed in–house to measure the proficiency of their new students. The Japanese Language Program (JLP) at International Christian University (ICU) is one institution that uses its own placement test.

Every September at ICU, the general JLP Placement Test (JLPPT) is administered to students in the JLP. The JLPPT consists of an aural comprehension subtest, a comprehensive structure subtest, and a reading *kanji* and vocabulary subtest. After three terms (Fall, Winter, and Spring), or nine months later, the same placement test is given to the same group of students to retest their proficiency. A report of the results of the September placement test, the retest in June, and the difference between the two tests (or gain score) is made to each student. For example, if students who were placed in the third level Japanese course (J3) in September on the basis of their JLPPT scores pass this course with a grade of A to D, they will successfully advance to J4 in Winter term and J5 in the Spring term. It is expected that the retest scores in June, at the end of the Spring term, will be equivalent to or at least close to the placement score necessary to get into J6. This should be the end–of–term proficiency score of a student who has mastered the J5 course requirements.

In reality, though, many students are disappointed to find that their retest scores are far below the scores which they expected. Why does this happen? Is there a similar effect for all students in all levels? In short, is there a mismatch between the standard required of those students who are placed into a level as compared to the proficiencies of those students who are promoted into the same level from lower courses?

Brown (1980) investigated proficiency data from two different student populations within some ESL classrooms and found that the two populations, newly placed students and continuing students (those promoted from lower courses), were significantly different. He also claimed that the difference in performance was expected to be greatest at the most advanced level (Brown, 1980, p. 113), but he did not investigate this last claim statistically. The results of Brown's study indicated that the newly placed students performed far better than continuing students on the three different measures: course grade, final examination, and cloze procedure. Brown suggested three possible causes: (a) the amount of instructional time devoted to the subject's ESL study, (b) the amount and nature of the subject's previous EFL study, (c) the amount of time that had passed since that previous EFL study (Brown, 1980, p. 116).

The present study focuses on issues similar to those investigated by Brown, i.e., the potential mismatch in proficiency between the students finishing a certain level with a passing grade in the Spring and going into the next level in the Fall and students newly placed in that level in the Fall. Unlike Brown's study, however, this study compares the proficiency score means of two different student groups in the same course level at different times (i.e., Fall and Spring terms) by using a pretest–posttest design.[2]

In the field of Japanese as a second language (JSL), there have been a number of studies about developing placement tests (Ichikawa & Ogawa, 1991; Ishida, Inagaki, & Nakamura, 1982; Kijima, 1988; Saegusa, 1986, 1988; Sakai, 1990), improving placement tests (Hiura, 1989; Suzuki, 1989; Taji, 1987, 1988), or measuring differences in the distribution of placement scores according to native language or other background variables (Sakai, 1988, 1990; Kano & Shimizu, 1991). However, to this researcher's knowledge, no studies have systematically compared and reported the placement test scores of different student populations across levels.

The purpose of the present study is to investigate potential proficiency differences between subjects in the Fall and Spring terms and possible reasons for such differences. If problems with placement are found, we should further study the appropriateness of decisions based on cut-point scores at each level, analyze the test items, and investigate the curriculum of the program. This study, then, formulates the following research questions:

1. Do continuing students in the Spring courses perform the same (at the end of their courses) as students at equivalent levels who are placed in the Fall (at the beginning of their courses)?

2. If there are differences in performance, are the differences observed at all levels or in some particular levels only?

Thus the following hypotheses are being studied:

H1: Newly placed students in the Fall will perform better than continuing students in the equivalent levels in the Spring.

H2: The differences between the two groups will be greater in the more advanced levels (following Brown's 1980 claim).

Method

Subjects

A total of 44 students from LEVEL4 to LEVEL7 took the JLPPT as a pretest (TEST1) and had scores on all subtests in the Fall semester 1992. This Fall semester group included all newly entering students and some students from a previous summer course.[3] In Spring semester, a total of 63 students from LEVEL4 to LEVEL7 took the JLPPT as a retest (TEST2) and had scores on all three subtests. The reason for choosing levels 4 to 7 will be explained in the *Procedures* section.

On the retest given at the end of each term, the scores are expected to be equivalent to the scores of students beginning the next level. Hence, the means at the beginning of each level (i.e., LEVEL4 to LEVEL7) on the Fall 1992 pretest (TEST1) were compared with the means at the end of the Spring 1993 retest (TEST2) for the level just below it (See Table 1). It was generally believed that the students taking the retest TEST2 at the end of one level would be equivalent to students just beginning the next level and taking the pretest TEST1.

Table 1. *Comparison of equivalent levels*

LEVEL	TEST1 (Beginning of course)	TEST2 (End of course)
LEVEL4	J4 (Beginning)	J3 (End)
LEVEL5	J5 (Beginning)	J4 (End)
LEVEL6	J6 (Beginning)	J5 (End)
LEVEL7	Advanced I (Beginning)	J6 (End)

The subjects in this study are described in terms of nationality, gender, academic status, and major for each level in Tables 2, 3, 4, and 5, respectively. Tables 2 through 5 indicate that the students on Test1 and Test2 (labeled T1 and T2) are similarly distributed in terms of nationality (Americans are predominant), gender, academic status (undergraduate is predominant), and major (language majors are predominant). Although these variables could become moderator variables in a study such as this one, the distributions appear to be approximately equivalent for the two groups so the potential effects appear to be controlled and will not be considered in this study. Note that such variables have remained uncontrolled in language studies elsewhere (Brown, 1988).

Materials

The only materials used in this study were the JLP Placement Test which was used for both the pretest (TEST1) and retest (TEST2). The separate scores on three subtests were used: the aural comprehension subtest, comprehensive structure subtest, and reading *kanji* and vocabulary subtest. Note that all of the questions on all three subtests were multiple-choice.

Basic descriptive statistics are given in Table 6. Notice in Table 6 that all three subtests are reasonably well-centered (as indicated by the means) and disperse the students well (as indicated by the relatively high standard deviations, or *S*). Notice also that the reliability estimates (both Alpha and K-R20) are reasonably high for the Aural Comprehension subtest and very high for the other two subtests.

Table 2. *Nationality*

	L3		L4		L5		L6	L7
Nationality	- T2	T1	T2	T1	T2	T1	T2	T1 -
Canada	2	1	1	1				
China				1			2	2
Denmark				1				
Germany	1			1			1	1
Ireland			2					
Japan				2		2		
Korea				2				2
Mexico							1	
Netherlands	3				2			
Russia							1	
Taiwan			1				1	
Turkey								1
UK	1	2	1				1	
UK/Hong Kong				1				1
USA	6	7	3	7	17	5	19	2
TOTAL	13	12	6	15	19	8	25	9

T1 = TEST1, T2 = TEST2

Table 3. *Gender distribution*

	L3		L4		L5		L6	L7
Gender	- T2	T1	T2	T1	T2	T1	T2	T1 -
Male	7	8	5	10	10	7	17	4
Female	6	4	1	5	9	1	8	5

Table 4. *Academic status*

	L3		L4		L5		L6	L7
Academic Status	- T2	T1	T2	T1	T2	T1	T2	T1 -
Undergraduate	12	8	6	11	18	8	18	7
Graduate	1	4		4	1		7	2

Table 5. *Major*

	L3		L4		L5		L6	L7
Major	- T2	T1	T2	T1	T2	T1	T2	T1 -
Language	5	5	5	7	13	6	14	5
Humanities	1		1		3		1	1
Social Science	6	4		7	3	1	8	2
Int'l Studies		1				1	1	1
Education	1		1	1			1	
Graduate Program				1				
TOTAL	13	12	6	15	19	8	25	9

Procedures

TEST1 was administered in the beginning of the Fall semester in 1992 as a general placement test. As described in the *Subjects* section, newly entering students and some students from the summer course took TEST1.

The students themselves selected which course series they would take: either the Intensive Japanese Course Series (twenty-two 70-minute periods a week) or the semi-

Table 6. *Placement test results (by subtest) for the two groups combined*

Subtests	k	Poss. score	N	Mean	SD	Reliability Alpha	Reliability K-R20
Aur. Comp	25	25	144	14.67	4.56	.7751	.7747
Structure	80	100	144	39.74	14.30	.9347	.9568
Reading	100	100	144	55.33	16.96	.9387	.9287

intensive course which is called the Japanese Series (ten 70–minute periods a week). Then they were placed in the appropriate level in either series according to their scores on the three subtests of TEST1. In this study, the mean scores of each subtest for the student groups taking Japanese Course Series J3 (beginning level), J4–J5–J6 (intermediate levels) and Advanced 1 will be analyzed. For simplicity, J3 to Advanced 1 levels will be labeled LEVEL3, LEVEL4, LEVEL5, LEVEL6, and LEVEL7, respectively. The scores of those students who were in the Intensive courses, J1, J2, and Advanced 2 of the Japanese series were not included in this study because those courses were not offered regularly and were missing in either the Fall or Spring semesters.

Analysis

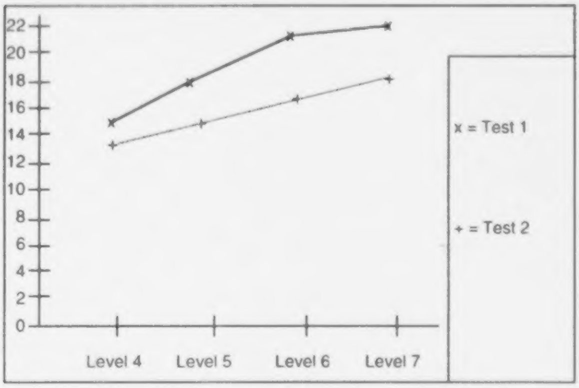
Essentially, the analyses in this study focus on the means for each level on TEST1 (administered as a pretest in Fall term) as compared with the means of students at equivalent levels on TEST2 (which was conducted program–wide at the end of the Spring semester in June, 1993).

In this pretest–posttest design, mean comparisons were first made by using two–way multivariate analysis of variance (MANOVA) procedures since there were three dependent variables (the three sets of scores on the three subtests). The two independent variables were the Tests (TEST1 and TEST2) and Levels (LEVEL4 to LEVEL7). Pillais, Hotelling and Wilks statistics were converted to F ratios to determine whether there were significant overall differences across the dependent variables for each factor (i.e., Tests and Levels). Then, where significant multivariate differences were found, univariate analyses were justified to discover where more specific significant differences might lie. Univariate analysis of variance procedures (and appropriate F ratios) were calculated to estimate the differences for Tests and Levels on the individual dependent variables, namely, the aural comprehension subtest (SCOREA), the comprehensive structure subtest (SCORES), and reading *kanji* and vocabulary subtest (SCORER). It is important to note that the subjects in this study were only those who had complete data (i.e., the results for those subjects with any missing scores were not

Table 7A. *Descriptive statistics and significance of differences in aural comprehension subtest (SCOREA)*

SCOREA	TEST1			TEST2			Mean differences
	n	Mean	S	n	Mean	S	(TEST2-TEST1)
LEVEL4	12	14.17	2.86	13	12.77	2.20	-1.40
LEVEL5	15	17.20	3.75	6	14.33	2.50	-2.87
LEVEL6	8	21.13	2.17	19	16.63	4.33	-4.50
LEVEL7	9	21.56	2.19	25	18.44	1.71	-3.12

Figure 1A. *Aural comprehension subtest (SCOREA)*



reported here and were not included in the analysis). Null hypotheses of no differences between group means were adopted and the alpha decision level was set at alpha < .01.

Results

The multivariate analysis of variance procedures (using Pillais, Hotelling, and Wilks tests) indicated significant overall differences for all three subtests taken together in Levels ($p < .001$; $df = 3, 99$) and Tests ($p < .001$; $df = 1, 99$), and showed no significant differences for the Levels by Tests interaction ($p < .01$; $df = 3, 99$). Univariate follow–up analyses for the main effects due to Levels and Tests indicated that significant differences were found ($p < .001$) for all three subtests on both variables. The descriptive statistics are given in Tables 7A, 7B, and 7C and the results are shown graphically in Figures 1A, 1B, and 1C.

Discussion

In short, all three subtests, SCOREA, SCORES, SCORER, were significant for Test and Level effects at $p < .001$, and none were significant (at $p < .01$) for the Levels by Tests interaction.

Figure 1B. *Comprehensive structure subtest (SCORES)*

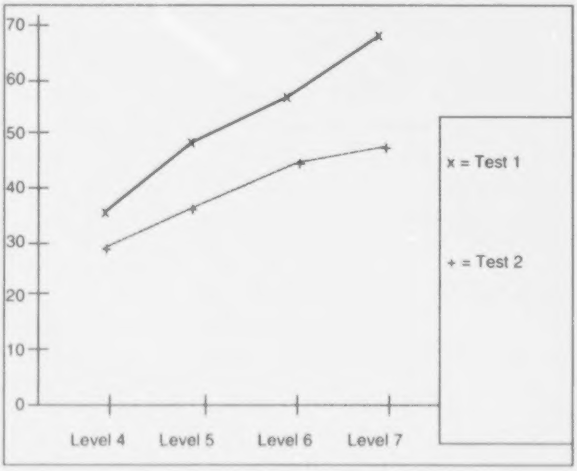


Table 7B. *Descriptive statistics and significance of differences in comprehensive structure subtest (SCORES)*

SCOREA	TEST1			TEST2			Mean differences
	n	Mean	S	n	Mean	S	(TEST2-TEST1)
LEVEL4	12	35.58	4.68	13	29.15	7.26	-6.41
LEVEL5	15	48.67	5.01	6	36.67	4.89	-12.00
LEVEL6	8	56.63	6.57	19	45.58	9.91	-11.05
LEVEL7	9	68.44	4.75	25	49.00	7.89	-19.44

Table 7C. *Descriptive statistics and significance of differences in reading, kanji, and vocabulary subtest (SCORER)*

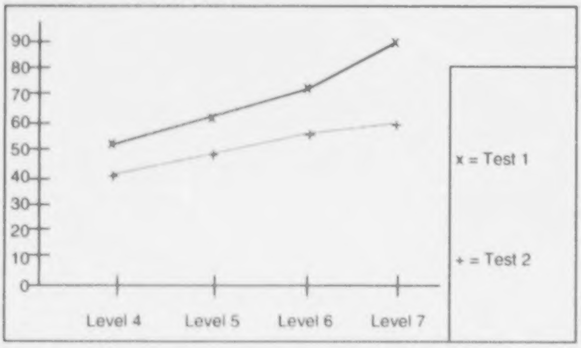
SCOREA	TEST1			TEST2			Mean differences
	n	Mean	S	n	Mean	S	(TEST2-TEST1)
LEVEL4	12	52.35	7.94	13	40.62	10.60	-11.73
LEVEL5	15	60.60	15.23	6	49.67	11.50	-10.93
LEVEL6	8	69.13	11.26	19	59.26	9.47	- 9.87
LEVEL7	9	87.89	4.35	25	64.60	10.99	-23.29

Answering the Research Questions

Research question 1 asked if continuing students in the Spring courses perform the same (at the end of their courses) as students at equivalent levels who are placed in the Fall (at the beginning of their courses). The results showed that placed students in Fall in all levels performed better. The findings supported the hypothesis (i.e., the newly placed students in the Fall will perform better than the continuing students in the equivalent levels in the Spring).

Research question 2 asked, if there are differences in performance, are the differences observed at all levels or in some particular levels only? In the SCOREA subtest, the differences become larger as the levels increase from L4 to L5 to L6), but the difference becomes smaller again when it gets to L7. For the SCORES and SCORER subtests, the differences are similar for levels L4, L5, and L6, but get larger for L7. However, since the MANOVA showed no significant interaction effect for Levels by Tests ($p < .01$), hypothesis 2 (i.e., Brown's hypothesis that the differences between the two groups would be greater in more advanced levels) is not supported by the results of this study.

Figure 1C. *Reading,* kanji, *& vocabulary subtest (SCORER)*



Other Pertinent Issues

Sample sizes. In this study, only complete data which fitted the research problem were used. The purpose of the study was to explore whether there were any differences between the student populations in the Fall and Spring semesters. Hence the sample sizes were relatively small.

Sampling procedures. Another problem existed with TEST2. Although the retest procedure (TEST2) was a program–wide procedure and was conducted in each classroom using instructional time, it was still up to the students to decide whether or not to take the test. Also, since advanced courses were offered separately (Aural Comprehension, Writing, and Reading), some students did not take some subtests, which caused some reduction in the data.

Also, since the subjects were all ICU exchange students, predominantly American, the generalizability of this study may be limited. Studies at other institutions with other types of students would be helpful in this regard.

Students' background knowledge. The length of Japanese language study, as well as teaching methods, and materials (including textbooks) in the previous institutions may have affected the relative performances of the groups taking TEST1 and TEST2. In other words, if the average length of previous Japanese study for students in any given level in the Fall was longer than that of students in the comparable level in the Spring, it is possible that the mean of TEST1 would be higher than the mean of TEST2 for those students due to differences in background alone rather than to differences in placed or continuing student status. Students' background should be taken into account in future studies.

Familiarity of the test. All the test questions were multiple–choice, and a computer–readable answer sheet was used for the purpose of computer processing the scores. This may have caused problems for some students. Some European and Asian students who are at intermediate and upper levels are not familiar with such a test format. They may have performed more poorly than their real proficiency level on TEST1 simply because of the novel answer sheet. However, this variable is very difficult to control.

Score distribution in each level. After taking the initial placement test, the students are placed according to the score ranges shown in Table 8 for each subtest. These norms were set based on the experience of the teachers who were involved in making each subtest. However, the norms were never checked statistically. Mismatches between the norms and the mean scores of the real levels of proficiency attained by the students may have caused the differences in the scores on TEST1 and TEST2. Closer study of this variable is necessary and important. However, again, such study is beyond the scope of this chapter, and the researcher only suggests that it should be done in the future.

Table 8. *Score ranges for placement using the JLPPT*

	Aural Comp.	Structure	Reading
LEVEL4	12 - 15	41 - 55	45 - 59
LEVEL5	16 - 18	56 - 65	60 - 69
LEVEL6	19 - 21	66 - 75	70 - 79
LEVEL7	22 - 23	76 - 85	80 - 84

Reliability and validity of the test. As shown in Table 6, all three subtests were reliable (Aural Comprehension K–R20 = .77; Comprehensive Structure K–R20 = .96; Reading Kanji and Vocabulary K–R20 = .93). However, the placement test was never studied to determine if it was correlated with other such tests (e.g., the *Test of Japanese Language Proficiency*) as a measure of criterion–related validity. Such a study and/or other content and construct validity studies would help in examining and improving the JLPPT.

Curriculum. In lower levels, course content is decided according to structural (grammatical) and functional syllabuses developed by the JLP. The placement test questions for lower levels were based on the new textbook which was recently developed by JLP. So, the students in the lower levels may have a better chance of achieving higher scores on TEST2 after they have studied this particular syllabus. In contrast, the syllabus for the intermediate levels and above is still under development by the JLP. Consequently, the curriculum for the upper levels is not firm and varies slightly in each term depending on which instructors are teaching in any particular term. Thus, even if the test questions are reliable, they do not necessarily reflect what the students at these levels were studying in their courses during the regular year. As a result, the intermediate and advanced students may have performed somewhat more poorly on TEST2. That is, the curriculum may have affected the results. However, this issue is also beyond the scope of this study. Nevertheless, further study of this question is needed.

Conclusions and Pedagogical Implications

By comparing the scores of the placement test administered in the beginning of the program with scores on the retest administered nine months later, this study revealed that there were potential differences between populations that should be similar at the beginning and end of the school year. The differences were consistent. There are several straightforward pedagogical implications of this study. First, the differences in the scores between TEST1 and TEST2 could be adjusted by applying more appropriate norms which reflect the scores of the population of each level on each subtest as they progress through the program. It may be important that such adjustments be made because the observed differences may discourage students who find, after taking the retest, that their scores were lower than expected after a year of study. This could cause a loss of interest or motivation in any student learning a language.

Second, the test questions for the intermediate and advanced levels should be carefully revised based on the syllabus taught in those levels so that they will reflect the content matter which is covered by the curriculum in each level. Of course, the purpose of a placement test is to spread students out into a normal distribution; and in that sense, it is a norm–referenced test. However, the researcher suggests that criterion–referenced type questions may also be added to reflect the "domain of language skills or material the students have learned" (as explained in Brown, 1988). In short, it would be helpful if the test were measuring what the newly entering students know and do not know about the syllabus of the particular institution in question.

This study on Japanese as a second language education found differences between newly placed students and continuing students just as Brown (1980) found in his ESL study at UCLA. The phenomenon may not be idiosyncratic to UCLA and ICU, but rather may be a more general trend at many language institutions. This phenomenon is not just an issue of test design, but also involves various topics such as curriculum design, students' background variables, etc. In fact, it is possible that effective testing that fits the curriculum is at the heart of any good language program.

Notes

[1] The Test of Japanese Language Proficiency (Society for Japanese Language Teaching) is administered once a year. The test consists of three subtests: aural comprehension, reading and writing (kanji), as well as grammar and vocabulary.

[2] The author sought and received permission to use the 1992 and 1993 placement test data from the Japanese Language Program at the International Christian University.

[3] In more detail, the Summer course (SCJ '92) students who took the JLPPT in September 1992 had already taken the same placement test at the beginning of SCJ '92 in July but wanted to see if they could skip a level. Students from SCJ '92 who did not wish to skip a level (i.e., wanted to continue to the next level in the Fall term with a passing grade) did not take this placement test.

References

Brown, J. D. (1980). Newly placed students versus continuing students: Comparing proficiency. In J. C. Fisher, M. A. Clarke, and J. Schacter (Eds.), *On TESOL '80 building bridges: Research and practice in teaching English as a second language* (pp. 111–119). Washington, DC: TESOL.

Brown, J. D. (1988). *Understanding research in second language learning*. Cambridge: Cambridge University Press.

Hiura, K. (1989). Placement test ni tsuite no houkoku [A report on the placement test]. *Bulletin of the ICU Summer Program in Japanese, 6*, 136–141.

Ichikawa, Y., & Ogawa, T. (1991). Kishuryoku shindan test no tameno bunpo map no ichishian [Grammatical map for a proficiency test]. *Annual Bulletin of Japanese Language Center. Tsukuba University, 7*, 155–176.

Ishida, T., Inagaki, S., & Nakamura, T. (1982). Hi–Kanjiken gakusei oyobi kikoku gakusei betsu nihongo nouryoku shindan houhouno kaihatsu. [Development of a proficiency test of the Japanese language for foreign students with non–Chinese background and returnees. *Monbusho kagaku kenkyuubi kenkyuu boukoku* [A study report for the Ministry of Education Scientific Research Fund] #551059.

Kano, C., & Shimizu, Y. (1991). Kanjiryoku no sokuei: Hyokani kansuru ichi shian. [Assessment of Kanji proficiency]. *Annual Bulletin of Japanese Language Center, Tsukuba University, 7*, 177–192.

Kijima, H. (1988). SPJ 1988 placement test houkoku [SPJ 1988 placement test report]. *Bulletin of the ICU Summer Program in Japanese, 5*, 109–114.

Saegusa, N. (1986). Development of item analysis and related programs for personal computer. *Annual Bulletin of Japanese Language Center, Tsukuba University, 2*, 193–200.

Saegusa, R. (1986). Placement test no toukeiteki shori no kokoromi [Statistical analysis of a placement test]. *Annual Bulletin of Japanese Language Center, Tsukuba University, 2*, 171–192.

Saegusa, R. (1988). Placement test no datousei to kongo no tenbo [Validity of a placement test and outlook for the future]. *Annual Bulletin of Japanese Language Center, Tsukuba University, 4*, 161–200.

Sakai, T. (1988). Placement test no bogobetsu bunseki. [An analysis of a placement test for students with different native languages]. *Annual Bulletin of Japanese Language Center, Tsukuba University, 4*, 139–160

Sakai, T. (1990). Placement test—Moji mondai ni kansuru ichi kousatsu [Placement testing—One view on writing]. *Annual Bulletin of Japanese Language Center, Tsukuba University, 6*, 167–186.

Suzuki, M. (1989). 1989 nen ICU kaki nihongo koza placement test ni kansuru houkoku [A report about the 1989 ICU Summer Program placement test]. *Bulletin of the ICU Summer Program in Japanese, 6*, 128–136.

Taji, Y. (1987). 1987 nen ICU Summer Program no Placement test houkoku—komoku bunseki no hitsuyousei ni tsuite [A report about the 1987 ICU Summer Program in Japanese placement test—the necessity of item analysis]. *Bulletinof the ICU Summer Program in Japanese, 4*, 120–121.

Taji, Y. (1988). Placement test komoku bunseki no houkoku [A report on the item analysis of a placement test]. *Bulletin of the ICU Summer Program in Japanese, 5*, 100–108.

Chapter 7

Evaluating Young EFL Learners: Problems and Solutions

R. Michael Bostwick
Katoh Gakuen

Assessment is one of several critical components of the instructional process because it lets the teacher know if the instruction has been effective and if the intended learning has occurred (Genesee & Upshur, in press). In essence, assessment answers the questions: How are we (students and teachers) doing? And, how can we do better? Unfortunately, far too few schools have systematic evaluations of their students' foreign language proficiency—evaluations that can provide the kind of information needed to improve and adjust classroom instruction. In particular, this lack seems to be especially prevalent in elementary school programs and conversation schools for young children.

Problems

Before selecting tests or designing their own instruments, teachers must be aware of three problems that can subvert assessment efforts in any English language program for children.

First, a lack of any clearly established program goals and objectives makes it very difficult to know what should be tested or if what is being tested is of any real importance (Winograd, 1994). Assessment is only effective and useful to the extent that the test content matches the primary goals of the class or program. Teachers may have fairly clear goals for an individual lesson but are often more vague about what they would like their students to be able to do six months, or six years from now. If your only goal was to teach the text and get to the end of the book, then your assessment would be fairly straightforward—just check to see if you got to the end of the textbook.

Unfortunately, there is no guarantee that the children will have acquired greater language proficiency in your march through the book. It is in just such cases that clearer proficiency goals can help you to maintain your focus where it should be: on developing the foreign language proficiency of the children. In short, assessment should always be tied to the intended goals of the language class or program involved because only with clearly understood goals can assessment be effective.

Second, teachers may not know exactly why they are assessing students, what type of information is desired, how they intend to use the information, and who the information is for. All of these questions are closely interrelated and inseparable. The answers to each must be clear in the mind of the teacher if assessment is to be of any value.

Third, a program may lack a clear and systematic plan for gathering information and assessing students' progress towards the school's goals and objectives. Without a plan, assessment becomes fragmented and disorganized, rendering it less effective and useful to all parties involved (Brown, 1992).

Traditionally, most formal English language assessment in Japan has been done to certify the level of knowledge of English and skill development so as to provide interested parties with information for selection (e.g., school entrance tests) or certification (e.g., Jido Eiken and Eiken STEP). The Eiken STEP test (from upper elementary students to adults) and the Jido Eiken test (for primary and upper elementary school students) are very popular in Japan. This is particularly true of the Eiken STEP test because the Japanese Ministry of Education (Mombusho) has authorized the test and given it their stamp of approval. Over 40,000,000 students have taken Eiken STEP in Japan over the last 32 years—a fact which attests to its popularity. The Eiken STEP claims to be a criterion-referenced test in that it specifies proficiency standards and attempts to identify whether the student can pass the pre-established standard. The test has seven different levels, and it is administered three times per year. The test is rewritten each time and a different version is used for each administration.

Unfortunately, the developers of the test do not publish information on the test's validity or reliability. This lack of information makes it difficult for people not in the Monbusho to know whether the test is doing what the publishers claim it can do--namely, identify students at distinct levels of language ability. Because there is no clear explanation of how levels or passing scores were determined and because no information on the reliability (consistency of the test from one administration to the next) is provided, we cannot be sure that the imaginary line indicating that a student has passed a certain level is not in fact jumping up and down with each different administration of the exam. In other words, it is possible that the same level test may be easier or more difficult from one administration to the next. The Jido Eiken for younger children has many of the same problems as the STEP test. In fact, the Jido STEP may be even more problematic because it is a newer test with less background and development.

The general trend of these tests has been in a positive direction with greater emphasis on communicative use and oral proficiency. Still, they are limited in the ways they can inform and influence daily classroom practices. One reason these tests may be of limited usefulness to teachers is that the goals and objectives assessed in these tests may not be aligned with the goals of a particular program. It is also true that, by the time you get the results from these tests, it may be too late to do much about it in your class because the students may have already completed the class or program. The tests described above are examples of summative forms of assessment in that they are generally used to provide a picture of students' overall achievement at the end of a course of study.

This type of summative assessment can be used to answer the following questions:

1. What level of achievement have the students reached?
2. Are the students progressing as well as expected?
3. How effective has the instruction been?

Formative assessment also provides feedback to teachers and students on the effectiveness of instruction and learning but in a much more diagnostic and prescriptive way (Herman, Aschbacher, & Winters, 1992). Formative types of assessment can render specific information on the progress of the students, what aspects of the instructional process need to be changed or reviewed, and what teachers may need to do next. Hence, formative assessment is generally more useful than summative assessment for improving teaching and learning.

Solutions

Teachers typically carry out formative types of assessment much of the time by monitoring learning and then adjusting instruction based upon informal observations and feedback from the students. However, if more formal and systematic assessment is to be carried out, the issues raised at the beginning of this chapter must be addressed. Only then can assessment serve its intended function.

Overcoming Problem 1: Have Program Goals and Objectives

As stated earlier, many English programs for children may lack clear goals or objectives other than the desire to complete the textbook. What exactly do you want the children to be able to do, know, or be like by the end of the class or program? In other words, what skills, knowledge, and attitudes would you like the students to be able to demonstrate as the result of being in your class? Defining clear goals and objectives helps to insure that what will be assessed is tied to the content of the course. Without clearly knowing where you are going and how you may get there, there is no guarantee that you may actually arrive.

One way to conceptualize this issue is by thinking of a continuum that ranges from goals that are very broad to goals that are extremely narrow (Oller, 1989). An example of a very general goal would be: Students will be able to understand and interpret written and spoken language on a variety of topics. An example of a very specific goal would be: Students can use the indefinite pronoun *other* in a sentence. The extremes presented here bring to mind a statement Albert Einstein is reported to have hung in his office, "Not everything that counts can be counted and not everything that can be counted counts." This certainly seems to be equally true in language testing as well (i.e., some objectives are important, others are not). The challenge for language programs is to identify those things that count and find a way to count them. Developing program objectives is no easy task because they must be at just the right level so that they identify meaningful, authentic, and important objectives while at the same time being specific enough so that they are observable, useful, and manageable for teachers. Once significant objectives are identified, programs must then ensure that the desired objectives actually drive assessment and instruction. It would make no sense to identify meaningful goals for a program and then continue to use the end-of-unit or end-of-book tests if these tests did not closely match the program goals.

The first step in designing effective assessment procedures is to be sure that you have well-articulated goals and objectives, without which assessment loses its usefulness. It could be argued that one of the main problems with the high school and university entrance examinations in Japan is (a) that there is no clear national consensus on what English language objectives should be and (b) that the assessments currently used by these schools have not been designed or aligned with such goals in mind. What then results is a tail-wagging-the-dog effect in which the assessments, not the goals, drive instruction.

Overcoming Problem 2: Know Who, What, How, and Why

Assessment can serve a number of different purposes and audiences, some of which may include:

1. Providing information to parents and students and address issues of accountability
2. Evaluating the strengths and weaknesses of a program and suggest areas for improvement
3. Identifying the placement of a student in a program
4. Providing diagnostic information about a child's language proficiency and skill development
5. Determining skill or proficiency mastery, and certify promotion or graduation requirements
6. Helping teachers focus their instruction and provide feedback to evaluate instruction
7. Providing students with feedback to help them see what needs to be done to achieve their goal and to encourage them to take more control over their own learning
8. Confirming student learning and enhance student motivation

Because assessment serves these many possible functions and audiences, it is important to be clear about exactly how the information is to be used and who it is for. Answers to these questions can greatly influence your assessment procedures (Genesee, 1994). It is also important to realize that assessment can be used to improve instruction and help students take control of their own learning. That is more easily accomplished when assessment is authentic and tied to the instructional goals of the program. Clearly then, assessment can serve purposes and audiences well beyond its traditional functions.

Some forms of assessment lend themselves more to certain kinds of purposes. For example, assessment that is for accountability purposes may be best done using summative forms of assessment (end of term tests, standardized tests, exhibitions, etc.). If the purpose is to inform instruction and to make decisions about the learning process, then observational checklists, anecdotal records, conferences, portfolios, etc. may be more helpful (Genishi & Haas Dyson, 1987).

Overcoming Problem 3: Have a Plan

Programs need to have clear and systematic plans for gathering information and assessing students progress toward the program goals if assessment is to be of any real benefit to the program (Brown, 1992). Without a plan, assessment becomes haphazard and disorganized, making it less effective and useful to all concerned. This is why assessment is integral to the learning process and must be planned before and during instruction. A plan should include: what will be assessed, how it will be assessed, when assessment will take place, and who will do the assessing. It should also include who the information is for, how the information will be used, and how the information will be reported.

Two Examples

There are many possible alternatives to traditional summative forms of assessment, several of which have been mentioned earlier. What follows are two specific examples of possible methods for collecting and documenting student progress. Both examples come from our English immersion program for elementary school children in which Japanese students study content area subjects in English on a daily basis. The purpose of both instruments is to: (a) establish accountability by providing a means to communicate progress to both students and parents, and (b) provide teachers and students with feedback on the effectiveness of the teaching and learning in the classroom. The content for each instrument is tied to the program goals and objectives as determined by both teachers and school administration.

Attainment Level Standards

English Attainment Level Standards can be designed to articulate overall program goals with specific benchmarks for each level (in our case for each grade) of the program. The general goals run across all grade levels and are broken down into (a) interpersonal, (b) educational, and (c) presentational communication goals. These overall goals are often difficult to measure because they are so global in nature; therefore, benchmarks that anchor these goals to observable behaviors are identified for each level. These benchmarks provide examples of behaviors/abilities one would expect from students at each level. Only two benchmarks for each goal at each level have been listed in the example given. The actual document contains many more benchmarks that could be incorporated into the instructional program.

Teachers and students work toward these specific benchmarks at each level and include time within the class for students to demonstrate proficiency with each of these benchmarks. Creating attainment level standards like the example provided in Figure 1 helps teachers monitor the progress of the students and provides on-going feedback that is integrated with the instructional program.

Figure 1 is an abbreviated version of the English Attainment Level Standards under development for our immersion program. However, even this abbreviated version reflects fairly ambitious goals. To be used in another context, it would need to be adapted to match the goals of that program and the needs of its students. Figure 1 is offered only as a resource to assist others as they begin to think about and develop their own attainment level standards. Schools may also wish to refer to other guidelines available when creating their own benchmarks. In our own case,

English Attainment Level Standards for Katoh School

Outcomes	Pre-1	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
1. Interpersonal *(Using language for non-academic purposes)* "Students engage in conversation, provide and obtain information, express feelings and emotions, and exchange personal opinions within informal, social contexts."	1A. Students can use expressions of greeting and leave taking appropriately ("See you tomorrow") and use set phrases with teacher and peers in and outside of the classroom. ("I'm finished" "I have a stomachache, etc.) 2. Students can understand and follow most teacher directions.	1A. Students can use classroom phrases to negotiate meaning and classroom conversations in English with the teacher 1B. Students can exchange essential information and ask and talk about personal likes and dislikes with each other	1A. Students can use classroom phrases to negotiate meaning and classroom conversations with each other in structured settings. 1B. Students can exchange essential information about events or transportation including date, time, and location	1A. Students can use classroom phrases to negotiate meaning and classroom conversations in English with each other in unstructured settings. 1B. Students can exchange information with peers about favorite activities and memorable past events.	1A. students can work in groups to plan events and activities to be carried out in English. They can evaluate the success of their task & identify ways to improve their communication. 1B. Students can exchange opinions and feelings about people and events in their personal lives.	1A. Students use appropriate social amenities: expressing gratefulness, extending & receiving invitations,apologizes, and communicating preferences 1B. Students can explain cultural habits and customs to foreign peers in English.	1A. Students can exchange information about current or past events and aspirations in their personal lives and the lives of their friends, family and community 1B. Students can exchange opinions and individual perspective on a variety of topics that are of contemporary or personal interest.
2. Educational *(Using language to learn & negotiate meaning in academic settings)* "Students discuss, question, hypothesize, explain solutions to problems, form opinions, make judgments in the second language on topics related to the school curriculum."	2A. Students can understand stories (picture books) and answer questions about the story with teacher guidance 2B. Students can describe 5 things about a picture of interest to the teacher.	2A. Students can respond to questions about a story and retell it in a coherent way maintaining a logical sequence 2B. Students can create stories or math word problems to tell other students.	2A. Students can give and follow simple instructions on how to play various games and ask questions if they do not understand. 2B. Students can paraphrase other students' ideas.	2A. Student can employ rephrasing and circumlocution to communicate messages successfully 2B. Students can recount main ideas in some detail and take part in structured, small group discussions	2A. Students can give and follow complex (4 or 5 step) directions to carry out a specific task and can ask questions for clarification. 2B. Students can work in groups to plan and conduct a math or science activity	2A. Student can exchange information and opinions on a variety of academic topics. 2B. Students can use persuasion and logical argument for the purpose of convincing them to see their point of views.	2A. Students use the dictionary thesaurus and other reference resources to select appropriate words for use in preparing and oral reports. 2B. Students can discuss and analyze the character plot development and themes found in authentic literature (poems, short stories, short works of fiction & non-fiction)
3. Presentational *(Using language to express personal meaning to others)* "Students present information, concepts, ideas, emotions, and opinions to an audience of listeners on a variety of scholastic topics."	3A. Students can draw a picture of their family (or other topic of interest) and ask and answer 2-3 questions from other students about the picture. 3B. Students can sing songs, do finger plays with teacher guidance	3A. Students can present simple oral reports / presentations about family members, friends, common objects in their environment. 3B. Students can recite memorized chunks of familiar stories and songs common to English speaking cultures	3A. Students can demonstrate / explain something of interest to their peers (show and tell) and give brief oral book reports. 3B. Students can create their own play based on a popular children's' story and perform it for others.	3A. Students can make a brief report on a personal experience. 3B. Students can summarize the plot and provide brief descriptions of characters in selected short stories and folk tales.	3A. Students can make a brief report to the class on a science or social studies topic. 3B. Students can perform short plays / stories / literature / songs in English.	3A. Students can prepare a video recorded message to peers abroad on topics of shared interest. 3B. Students can describe / present some aspect of Japanese life or culture to peers of the target culture.	3A. Students can summarize the contents of a feature magazine / newspaper article on a topic of current or historical interest to peers of the target culture. 3B. Students can debate a topic of interest, presenting and defending their position logically and persuasively.

Figure 1. Attainment Level Standards

BEST COPY AVAiLABLE

both the ALL Guidelines (Scarino, Vale, McKay, & Clark, 1988) and ACTFL (1995) national standards were consulted in developing our own framework. Other schools may also find these resources helpful.

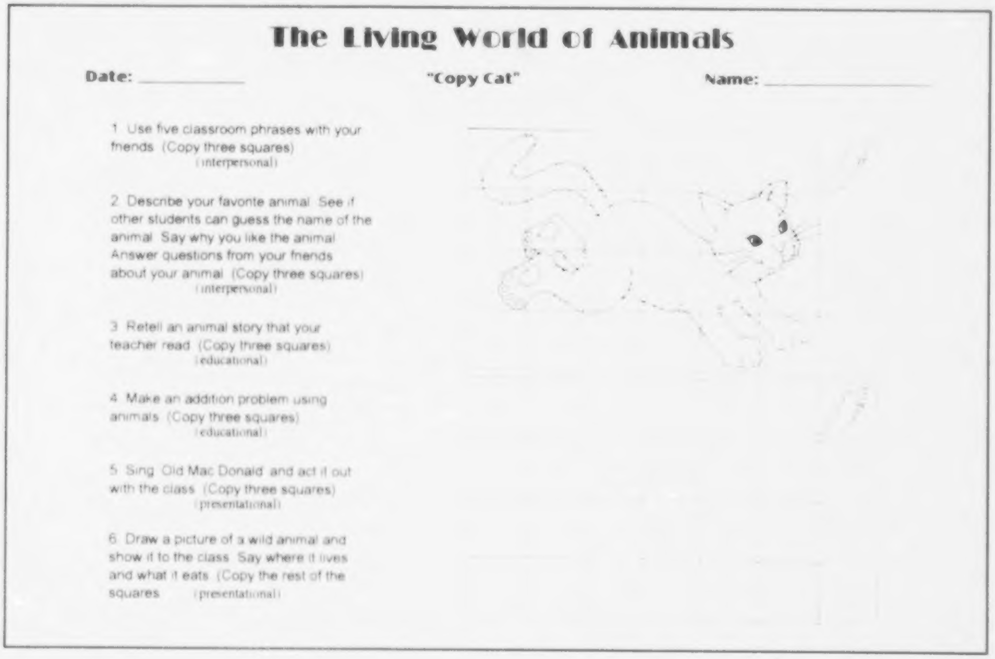
Unit Checklists

The Unit Checklists (sometimes referred to by our students in the past as *star sheets* because the teacher places a star next to each item mastered) are directly related to the objectives for each 3 to 6 week unit of instruction in the program. Each checklist contains approximately six tasks that students are expected to be able to do by the end of the unit. In a sense, they are mini-exhibitions of student achievement within each unit because the successful exhibition of these tasks demonstrates student mastery of the objectives of the unit.

Having clearly established objectives for each unit—objectives that are written and given to both students and teachers at the beginning of the unit—helps to provide direction to the learning and ensures that assessment, instruction, and objectives are aligned with each other. Because the objectives are achievable within a relatively short period of time, students have a clearer understanding of their responsibilities in the classroom and are motivated by successful accomplishment of each objective. Information gained from these activities provides direct feedback to students as to how they are progressing, i.e., objective evidence that they are learning. However, by far the greatest power of these checklists is that they clarify for the students what is believed by their teachers to be the important objectives of each unit.

Figure 2 is an example of a primary level unit on animals with six tasks. The tasks act as benchmarks of student achievement and assist teachers and students in documenting mastery of course objectives in an on-going formative manner. Although tasks normally listed at the unit level may be more detailed than many of the benchmarks identified in the Attainment Level Standards mentioned above, some of the

Figure 2. *On animals with six tasks*



The Living World of Animals

Date: ____________ **"Copy Cat"** **Name:** ____________________

1 Use five classroom phrases with your friends (Copy three squares) (interpersonal)

2 Describe your favorite animal. See if other students can guess the name of the animal. Say why you like the animal. Answer questions from your friends about your animal. (Copy three squares) (interpersonal)

3 Retell an animal story that your teacher read. (Copy three squares) (educational)

4 Make an addition problem using animals. (Copy three squares) (educational)

5 Sing Old Mac Donald and act it out with the class. (Copy three squares) (presentational)

6 Draw a picture of a wild animal and show it to the class. Say where it lives and what it eats. (Copy the rest of the squares (presentational)

63

BEST COPY AVAILABLE

tasks (benchmarks) on any given *star sheet* may actually come from this document.

In the example in Figure 2, each task is tied to the overall language goals of the program (interpersonal communication goal, educational communication goal, and presentational communication goal, as shown in Figure 1). Because immersion teachers are also responsible for teaching the content areas in the elementary school curriculum, the tasks also reflect academic goals as well. The unit checklist provides two tasks for each of the three overall language goals of the program. Teachers check off students as they demonstrate their ability to actually do the task successfully. These records are kept and can later be shared with parents.

Conclusion

I have argued that assessment should ultimately be tied to the goals and objectives of the program and the needs of the students. Assessment should also reflect the purpose you have for collecting the information and it should be done in a systematic way. In addition, assessment can serve several functions. It can inform teachers about the extent to which their students have learned what they set out to teach. But perhaps at an even more important level, assessment can be an integral part of the instructional process. Since effective assessment enables teachers to know how successful their instruction has been, the information collected should be used to inform and influence on-going classroom practice.

In short, assessment is an essential part of the learning process and not something added on at the end of a series of lessons (Brown, 1995). The two examples provided here demonstrate how assessment that (a) is aligned with the program goals, (b) is clear about its function and audience, and (c) has a systematic plan for gathering and reporting the information, can provide teachers and program managers with the kinds of information needed to improve instructional practice in their classrooms.

References

Brown, J. D. (1992). *Testing in language programs.* Tokyo: Temple University Japan.

Brown, J. D. (1995). *The elements of language curriculum.* Boston, MA: Heinle & Heinle.

Eiken, STEP. (1995). *Eiken, STEP.* Tokyo: Nihon Eigo Kentei Kyokai.

Genesee, F. (Ed.). (1994). *Educating second language children.* Cambridge: Cambridge University.

Genesee, F., & Upshur, J. (in press). *Alternatives in second language assessment.* Cambridge: Cambridge University.

Genishi, C., & Haas Dyson, A. (1987). *Language assessment in the early years.* Norwood, MA: Ablez.

Herman, J., Aschbacher, P., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Jido Eiken. (1995). *Jido Eiken.* Tokyo: Nihon Jido Eigo Shinko Kyokai (JAPEC).

ACTFL. (1995). *National standards in foreign language Education.* Washington, DC: American Council on the Teaching of Foreign Languages.

Oller, J. W., Jr. (1989). Testing and elementary school foreign language programs. In K. Muller (Ed.), *Languages in elementary schools.* New York: The American Forum.

Scarino, A., Vale, D., McKay, P., & Clark, J. (1988). *ALL guidelines.* Melbourne, Australia: Curriculum Corporation.

Winograd, P. (1994). Developing alternative assessments: Six problems worth solving. *The Reading Teacher, 47,* 420-423.

Section III

Standardized Testing

Chapter 8

Good and Bad Uses of TOEIC by Japanese Companies[1]

MARSHALL CHILDS
FUJI PHOENIX COLLEGE

DIRECTIONS: This is a quiz. Read the following passage and answer the questions.

Passage

You are the Education Director of a Japanese company. You are in charge of the first-year training of a group of 113 new employees, recruited from college. Their average TOEIC score upon entry into your company was 269, thanks to the Japanese education system. Your program gave them 53 hours of classroom English during the course of one year. You have measured their English-language proficiency gains by means of TOEIC tests administered before and after the English instruction.

Now it is time for you to look at the TOEIC results, draw conclusions, and take action. You review what you know about the uses of TOEIC. For each of the following uses of TOEIC, write either **Good**, **Bad**, or **Good Under Some Circumstances** (GUSC).

Questions

1. ___________ Measuring overall group gains in proficiency.
2. ___________ Comparing the performance of different schools or treatments.
3. ___________ Gauging the progress of individual learners.
4. ___________ Counseling learners on their progress.
5. ___________ Guiding the courses of study of individual learners.

Readers may be pardoned if they do not know the answers to all these questions. Most company education directors do not know the answers, either. This chapter will give the answers in the course of discussing the issues, but for readers who require immediate resolution of ambiguity, an answer key is offered in the *Conclusion* section of this chapter.

The *Test of English for International Communication* (TOEIC) is both a boon and a bane for company education directors. It is a boon in that it is widely available, internationally understood, and relatively cheap to administer. It is a bane in that TOEIC is too easy to use for purposes that it should not serve, a circumstance that leads to ineffectiveness and in some cases harmful misunderstandings. One problem is a general lack of understanding of the difference between *norm-referenced tests* and *criterion-referenced tests* (see, for instance, Bachman, 1990;

Brown, 1995; or Chapter 1 of this book). TOEIC is a norm-referenced test, which is to say that its chief virtue lies in its ability to spread the scores of test-takers into a bell-shaped curve of all test-takers. TOEIC is not a criterion-referenced test, which is to say that it is not designed to measure progress toward mastery of particular instructional objectives in the way that, for instance, final examinations measure the degree to which learners have mastered the subject matter of formal instruction.

In an ideal world, each type of test—norm-referenced and criterion-referenced—would be used in its own sphere, without overlap or confusion. In fact, however, tests are sometimes used inappropriately. In particular, the use of TOEIC in Japan often seems to be an attempt to combine the functions of norm-referenced and criterion-referenced tests. Although TOEIC is basically a norm-referenced test, the sellers of TOEIC themselves encourage users to regard the test as a measure of progress of English-language students (see, for example, applications described approvingly in the quarterly Japanese-language publication, *TOEIC Newsletter*). As a result of the above factors, many companies, governments, and educational institutions use TOEIC to measure both English ability level and language gains due to teaching/learning.

The Present Study

This chapter describes in greater detail the situation presented in the short quiz at the beginning, in which a Japanese company requires recently hired employees to take a series of TOEICs, not only to measure their overall English skill but also to measure their learning progress in groups and as individuals, and to guide their individual courses of study. Because the company employs different language schools for teaching, it also compares the schools on the basis of the TOEIC gains that occur under their tutelage. The purpose of this chapter is to examine the uses of TOEIC by this company in an effort to judge what are good and bad uses of the test results.

Questions of interest include the following: Is TOEIC a good measure of the average gain for a group of learners? Is an apparent difference between the average gains of two subgroups significant and meaningful? In a series of four TOEIC tests in which the final test does not show the highest average score, how is learning gain to be measured? How are we to understand an overall *decline* in mean scores after training? Is it possible to measure and interpret the standard error of measurement of the learners in this study? How can we understand the levels of individual learners when, in a period of generally increasing English proficiency, ups and downs introduce uncertainty into the interpretation of test scores? Given this uncertainty, how can we assign learners to ability groups for further English-language training and for job selection? How shall we counsel individual learners on their apparent progress or lack of it?

Method

Subjects

Subjects consisted of 113 recent college graduates (110 men and three women) who joined a Japanese manufacturing company as new employees in April 1992. All were native speakers of Japanese. Precise age data are not available, but it is certain that most subjects were 22, 23, or 24 years old during the course of the study. During the period of observation, English-language training was not the primary goal of the subjects. They were engaged in a grueling one-year initiation to the company that included many practical experiences as well as classes on various subjects. Outside the English classes, they rarely used English on the job.

The mean TOEIC starting score of this group of learners was 269. This mean starting score is approximately 100 points lower than the average scores of all recent college graduates in Japan according to the Vice-Chairman of IIBC, the official sponsor of TOEIC in Japan (Kitaoka, 1992). The difference may be due, in greater or lesser part, to three causes: (a) company hiring practices, (b) the fact that test-takers were mostly men, and (c) the fact that the company tends to assign beginning employees who achieve unusually high scores to a different course of English training, more advanced than the courses described here. A skewness statistic was applied in order to judge whether the initial scores of these employees formed a normal distribution. They did.

Materials

Testing of English-language ability was done by means of the TOEIC, a two-hour test of EFL listening and reading (TOEIC Steering Committee, 1991). This test was originally designed to:

1. spread out the scale of the TOEFL test, particularly in the lower half of that scale,
2. provide "highly valid and reliable measures of real-life reading and listening skills" as needed in business and government (Woodford, 1980, p. 5), and
3. offer means of score interpretation "that would allow score recipients to actually see the kind of English that examinees at different levels could read" (Woodford, 1980, p. 5).

Procedures

Learners were given a total of 53 hours of formal English teaching, consisting of (a) five-day (37-hour) intensive classes conducted in groups of 10 to 12 students during September and October 1992, and (b) four monthly half-day (4-hour) English classes conducted in November and December 1992, as well as January and February 1993. All 113 learners received the treatment; there was no control group. Sixty-five of the subjects began their training in Tokyo and were transferred to Osaka in late December. Forty-eight began in Osaka and transferred to Tokyo in late November.

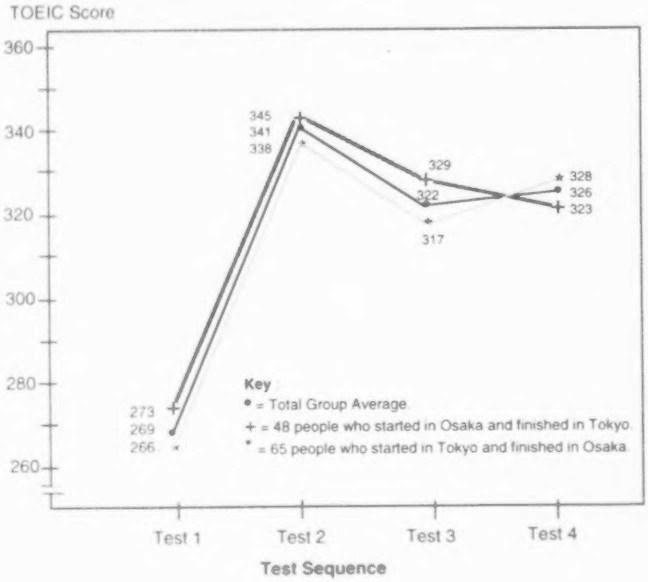
The TOEIC was administered on four occasions: at the time of hiring in April 1992; at the end of the five-day intensive English courses in September and October 1992; and at the end of the second and fourth monthly half-day classes, i.e., in December 1992 and February 1993. These administrations will be referred to below as test 1, test 2, test 3, and test 4. Tests were administered by TOEIC-accredited test centers in Tokyo and Osaka.

Results

Group Means

Figure 1 shows the overall shape of the mean TOEIC results for all 113 learners and for the two subgroups who changed places between Osaka and Tokyo during the program. The major point is that overall scores increased. It is notable, however, that the highest average scores were recorded in test 2, at the end of the one-week intensive program for all groups. Also, the subgroup that

Figure 1 *Mean TOEIC test results for 113 learners (total group and two subgroups that finished in Osaka and Tokyo, respectively)*



began in Tokyo and finished in Osaka appears to have outperformed the overall mean, beginning below the mean and finishing above it.

Overall mean scores increased from 269 to 341 from test 1 to test 2. Overall mean scores for tests 3 and 4 were 322 and 326, respectively. A repeated-measures one-way ANOVA showed a significant difference among these means, ($F = 61.12$; df = 3, 336; $p < .05$). Follow-up Scheffé analyses for $p < .05$ showed that the means of tests 2, 3, and 4 were all significantly different from the mean of test 1, and also that the mean of tests 3 and 4 taken together was significantly different from the mean of test 2, but that the means of tests 3 and 4 taken individually were not significantly different from test 2 or from each other. To say that a relationship is "statistically significant" means that the relationship is probably *not* due to chance alone (Brown, 1988, p. 122).

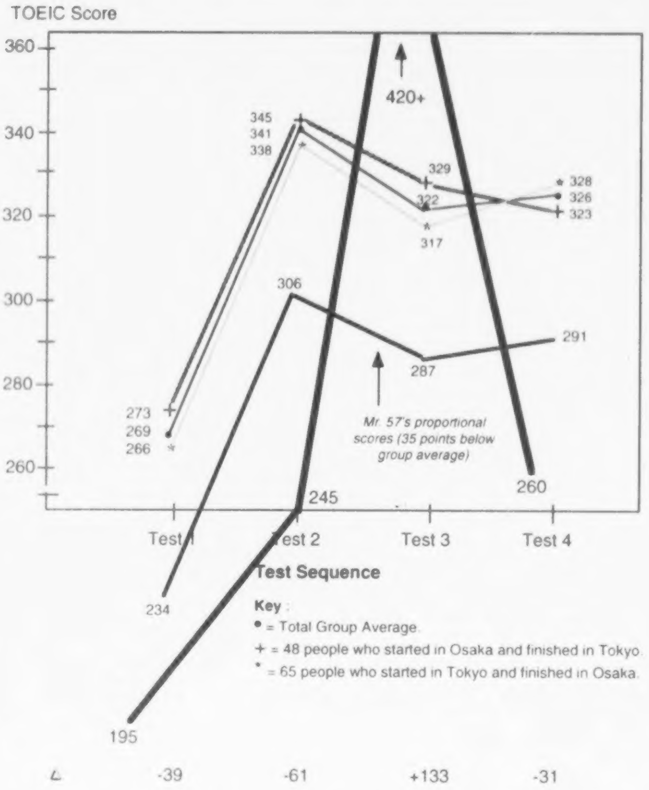


Figure 2. *Scores for learner #57: Actual scores, calculated proportional scores, and the differences between them*

Figure 1 shows that the Osaka and Tokyo group means follow the general pattern of the overall means. The group that began in Tokyo and finished in Osaka started with a mean lower than the overall mean (266 compared to 269) and finished with a mean higher than the overall mean (328 compared to 326). The group that began in Osaka and finished in Tokyo began higher (273 compared to 269) and finished lower (323 compared to 326). These differences were not found to be statistically significant.

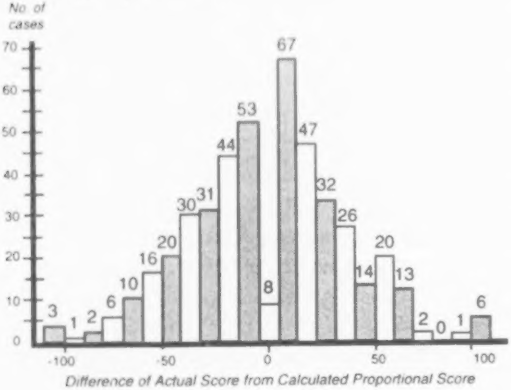
Reliability

To assess the reliability of TOEIC in this setting, three procedures were used: an internal consistency reliability estimate, a standard error of measurement estimate, and an examination of the differences between actual scores and calculated proportional scores. To assess internal consistency reliability, the Kuder-Richardson formula 21 (K-R21) was applied to the scores for all four tests, using raw scores approximated from the standardized scores by reversing the procedure of score conversion (ETS, 1980).[2] Approximated raw scores ranged from 53 to 110 (out of a possible 200). The resulting K-R21 reliability estimate was .57. This means that 57% of the variance of test scores was consistent on the test, and 43% was from other sources. If it is in error, K-R21 tends to underestimate internal consistency reliability (Guilford & Fruchter, 1978, p. 429). Nonetheless, .57 is not high. Accordingly, an examination of the standard error of estimate will be of interest.

The standard error of measurement (SEM) is a gauge of the extent to which an individual's actual score varies from his or her "true score" (the "true score" is defined as the score that a learner would achieve if he took the test a very large number of times without learning anything). With the present data, the SEM was calculated by the formula **SEM = SD 1-R**, where SEM is standard error of estimate, SD is standard deviation, and **R** is internal consistency reliability as measured by K-R21. Based on this formula, the SEM for the students in this study turned out to be seven test questions, or 43 points in terms of the converted TOEIC score. This SEM statistic can be interpreted to mean that there is a 68 percent chance that a given test score is within 43 points of the learner's "true" score. It follows, of course, that there is a complementary 32 percent chance that a given test score differs from the true score by more than 43 points. Some consequences of this SEM will be discussed below.

The fact that these learners took four tests in sequence permits us to gain some idea of the consequences of measurement error. If there were no measurement error, we would expect each learner to maintain his or her proportional distance from the group average on each of the four tests. The learner with ID #57, for instance, averaged 35 points below the group means. Recall that the overall group means for the four tests were 269, 341, 322, and 326. Accordingly, we might expect that Mr. 57 would have scored 35 points below the mean in each case, chalking up scores of 234, 306, 287, and 291 (see Figure 2). Of course his actual scores were not so beautifully proportional; instead, they were 195, 245, 420, and 260 (Figure 2). They differed from the calculated proportional scores by, respectively, -40, -62, +133, and -31 points.

Figure 3. *Differences between actual and calculated proportional scores for 452 TOEIC scores (grouped data)*



Such differences between actual observed scores and proportional scores were calculated for each of the 113 learners in the present study. A graph of grouped data is shown in Figure 3. Figure 3 is, in effect, a graph of the distribution of the standard error of measurement (SEM). It will be seen that the graph resembles a normal curve. The distribution may be further described as follows: differences within plus or minus 35 points make up 67.9% of the total, differences within plus or minus 70 points make up 95.4% of the total, and differences within plus or minus 105 points make up 98.7% of the total. Thus the distribution of differences from proportional scores is similar to that of a normal distribution with a standard deviation somewhat greater than 35. This distribution of differences is uniform in the sense that it is the same for learners at all levels: for the top third, the middle third, and the bottom third of the group of 113 learners, the distributions of differences all show the same form.

Using the present data to examine patterns of ups and downs, learner experiences were classified in terms of sequences of score increases and decreases for tests 1, 3, and 4. Test 2 results were omitted from this analysis because they appear to include an uncharacteristic "polish" effect (see the third part of the Discussion section) not found in the other three sets of test results. Table 1 shows the number of learners who experienced each of the four possible sequences of *gain* and *no gain* in the transitions from test 1 to test 3 and from test 3 to test 4. In a period when the mean score increased by 56 points, only 35 learners (31%) experienced two successive gains. Seventy-five learners (66%) had mixed experiences and three learners (3%) had two successive losses.

Discussion

What lessons for TOEIC users can be gleaned from the results of this analysis of the scores of 113 learners? We have seen that some group mean results are statistically significant and some are not. Using a standard internal consistency reliability estimate, TOEIC does not appear to be very reliable, at least for this group of students. And we have seen that the standard error of measurement is large for this group of students. Some further discussion of each of these areas, along with some related issues, is warranted. The following discussion will cover six topics: (a) measuring overall group gains, (b) facing ambiguity about the causes of gains, (c) assessing temporary versus long-term gains, (d) using test results to evaluate different teaching treatments, (e) gauging the learning of individuals and basing action upon it, and (f) counseling and guiding individual learners in the face of ups and downs in test scores.

Measuring Overall Group Gains

In the opening section of this chapter, the question was raised: "Is TOEIC a good measure of the average gain for a group of learners?" Based on the present data, the answer is that, with appropriate statistical tools, significant gains can be discerned in the means of large groups. The data on 113 Japanese learners from the company in this study showed a significant gain, in the sense that for the overall group of 113 learners, tests 2, 3, and 4 were significantly different from test 1. One caution should be observed, however: the *causes* of the gain have not been demonstrated. The gain, or part of it, may well be due to formal classroom teaching, but without further information, alternative explanations of the improved performance cannot be excluded. This is a serious handicap in interpreting the results of this training and testing program. Some alternative explanations for the learning gain are explored in the following section.

Table 1. *Patterns of gain and no gain for 11 test-takers in a sequence of three TOEIC tests*

Test 1 to Test 3	Test 3 to Test 4	Number of Learners
Pattern		
Gain	Gain	35
Gain	No gain	58
No gain	Gain	17
No gain	No gain	3
Total number of learners		113

Causes of Score Gains

In the present study, it is difficult to separate the effect of teaching from other possible causes of score gains. Four major causes may be at work: formal English teaching, differences in motivation, English in the environment, and practice effect. Each of these possible causes will be considered in turn.

The formal teaching of English. Certainly this is the first cause to look at. Formal teaching seems unlikely, however, to account for the whole gain. There were only 53 hours of formal teaching, and a gain of 56 points, or about one TOEIC point per hour of teaching. According to Saegusa (1983, 1985), in the TOEIC range of around 730 a gain of one TOEIC point would require about 2.5 hours of teaching. The ETS (1990) observes that the difficulty of achieving a TOEIC score increases out of proportion with the level of the score:

> . . . it is much easier for an employee to improve a score at the lower end of the scale than at the upper end. The employee with a base score of 240 can improve 150 points in a much shorter time than, for example, the individual whose score is 710. (p. 6)

Although the 113 learners are generally in the TOEIC range of 250 to 400, a teaching efficiency of one TOEIC point per teaching hour would be a difference of 2.5 times from the range of the low 700s. A difference of this magnitude from one part of the scale to another seems rather large.[3] It is probable that the observed learning gain of 56 points is due to a combination of effects, of which teaching is only one.

Differences in motivation levels. The lower results on test 1, compared to the later tests, may be due in part to a lack of focused motivation before the learners had actually joined the company where they now plan to do their life's work. In later months the learners may have become more motivated. Hence, there may be a noticeable difference in motivation between the first and subsequent tests.

Heeding English in the environment. For a 22-year-old college graduate, living in a major Japanese city for one year may by itself enhance proficiency in English. Working for a major company, watching television, listening to the radio, and reading advertisements in cosmopolitan Japan all involve exposure to English, and young professional people might glean some knowledge of that language from their environment.

Practice effect. It seems likely that a learner who takes the TOEIC several times will become more proficient at taking the TOEIC. For learners who learn greater test-taking skills, some portion of any score increases may be due to this practice effect, rather than to any increased proficiency in English.

Temporary versus Long-Term Gains (the Polish Effect). The fact that the highest scores were achieved in the second test, rather than the third or fourth, deserves comment. The second test was administered at the end of a one-week full-time course in English. After that, the learners returned to their predominantly Japanese environments and experienced only one four-hour English class per month. Thus the high scores for the second test may be attributed to the active practice that immediately preceded that test administration. This effect, which is longer than short-term memory but which nevertheless degrades over time, will be called the "polish" effect here, in reference to the fact that at the end of an intensive week, the learners' English was highly polished. When active participation in English became less frequent, the polish gradually lost its luster, leaving only the underlying longer-term gains to be reflected in the results for tests 3 and 4.

The mean scores for tests 3 and 4 are still significantly higher than the starting results (test 1) but indicate retention of only about two-thirds of the polished gains shown in test 2. This fraction (two-thirds) may be a good rule of thumb for estimating the amount of learning that is retained after the polish of an intensive class fades. Further research may be rewarding, however, not only to check the magnitude of the polish effect under different conditions, but also to explore the effectiveness of different approaches to scheduling English classes. It is possible, for instance, that even after subtracting the loss due to the polish effect, teaching English intensively in large blocks will prove to be more effective than presenting it in small, diluted doses.

Comparing Teaching Treatments

The company hires commercial schools to teach English to its employees, and uses mean TOEIC gains of learners taught by the different schools to compare the schools' performances. Such a comparison is possible, with due caution. In the present data, we have seen that, using ANOVA statistical procedures, mean score differences of about 50 points for large groups of learners can be statistically significant when means and standard deviations are similar to those found here. However, in the present data, the differences in the means of the Osaka and Tokyo subgroups were not statistically significant. This may be taken as an indication of the need for caution—and for proper statistical procedures—in using TOEIC to assess differences in the performances of subgroups.

If the different gains of the Tokyo and Osaka subgroups are not statistically significant, how should we think of them? The differences in the means may still be of interest not only to education directors but also to school administrators as early warnings. Thus, the director of the Tokyo school might look upon these results as an occasion for re-examining the school's practices in an effort to maximize the school's teaching effectiveness.

In view of the discussion in the previous sections, two primary cautions should be observed in comparing any mean gains for subgroups. First, causes are almost always ambiguous, and should be considered as potential sources of doubt for any conclusions. Even if we can identify differences between two groups of learners, we cannot know how much of any reported gain is due to teaching. Second, as ETS (1990) says, it is more difficult for higher-level learners than for lower-level learners to gain a given amount. Thus, if the learners' beginning levels are different, and particularly if assignment to schools is based on level of proficiency, differences in TOEIC score gains should be interpreted very carefully. If it is really true, for instance, that for students in the 700 range it takes 2.5 times as many teaching hours as in the 240 range to raise a TOEIC score by one point, then a one-for-one comparison of mean score gains would be inappropriate. With these caveats in mind, administrators using proper statistical procedures can examine the differences in mean scores of learners in different schools or of learners given different treatments in in-house education programs.

Gauging Individual Learning

To summarize the findings on reliability, we have seen that the SEM was 45 points for the subjects in this study, and we have also seen (Figure 2) that in the sequence of tests these learners' TOEIC scores fell within 35 points of their calculated proportional scores only about two-thirds of the time. In addition, we have seen (Table 1) that, in a group whose overall mean score was increasing significantly, two-thirds of learners saw no gain in their scores in one or the other of two comparisons.

In light of these results, TOEIC seems a poor instrument for gauging the short-term learning gains of individuals like those in this study. Hence in a situation like ours, TOEIC should not be used to assess the progress of individuals in programs with relatively few teaching hours, or should be used only with extreme care. For gauging the effects on the individual of a teaching program of considerable length (perhaps in the range of 200 classroom hours), or for measuring very long-term growth of English ability (measured in years rather than months), TOEIC may be used with caution as a very general measure of change in proficiency. The caution to be born in mind is, of course, that scores are subject to a standard error of measurement that may be in the range of the apparent proficiency gain.

The fact is simply that TOEIC, as a norm-referenced test, is not the best gauge of individual learning. Instead, criterion-referenced tests, which are specifically designed to measure individual learning, should be used. A well-designed criterion-referenced test measures the learner's degree of mastery of the specific material being taught. When applied as a pretest and a posttest, a criterion-referenced test can be a very reliable measure of learning gain. See Chapters 1, 2, and 3 for information on creating and using criterion-referenced tests.

Counseling and Guiding Learners

As Table 1 illustrates, in a sequence of four tests, individual learners can expect to see some inexplicable, and sometimes very disappointing, differences in their scores. The ups and downs make it difficult also for administrators to use TOEIC scores for counseling individuals on their progress and for prescribing specific instructional levels or experiences.

How should learners be counseled about the vicissitudes of their TOEIC scores? Probably the first message to give them is that time and chance happen to them all: that jumping around is in the nature of TOEIC scores. Showing learners Figure 2 or a similar diagram should help them understand the probabilistic relation between TOEIC scores and underlying proficiency. A second message is that TOEIC scores will very likely reflect large differences in learning; TOEIC score gains are unreliable only for differences that run afoul of the standard error of estimate. A third message is that TOEIC is not designed to provide information on specific strengths and weaknesses or on specific language topics that the learner should address.

For these reasons, TOEIC should *not* be used except in a very general way for placement decisions, selection of teaching treatments, and counseling of learners. Criterion-referenced tests should be used for other purposes like diagnostic testing, assessing achievement, or measuring learning gains. Program administrators should realize, too, that the learners share the responsibility for identifying strategic targets for their own learning. When that responsibility is taken seriously, the learner will seek out activities that lead to unmistakable improvements in TOEIC scores. The wise administrator will tap the energy and knowledge of the learners in selecting learning experiences. Together, the administrators and teachers with their general knowledge and the learners with their specific knowledge can set a better course than either group can do alone.

There is a positive aspect to the variability of TOEIC scores. Students may be counseled that if they take the test several times, they can expect that by chance alone they will achieve a score that is higher than their true score. This is a positive motivation for learners, for the TOEIC score that is kept on record is the highest score, not an average of scores. A sporting appreciation of the odds may encourage learners, who face repeated TOEIC tests, to attack the tests with vigor.

Conclusions

We are now in a position to answer the questions in the quiz at the beginning of this chapter. How good is TOEIC for the following uses?

1. Measuring overall group gains in proficiency: *Good under some circumstances.* We have seen that TOEIC can be used effectively to differentiate some group mean gains in a group of 113 learners, with the caveats that careful measurement of statistical significance is necessary in order to distinguish real gains from illusory ones, and that, even if the significance of gains is established, the causes of gains may remain problematic. In addition, the administrator must consider to what extent any differences measured by TOEIC are related to what has been taught.

2. Comparing the performance of different schools or treatments: *Good under some circumstances.* The cautions of answer 1 apply here. In addition, the administrator must be aware that the difficulty of raising a TOEIC score is considerably greater at the upper end of the scale than at the lower end. Therefore, TOEIC should be used with care for measuring the effectiveness of different teaching approaches.

3. Gauging the progress of individual learners. *Bad.* The use of TOEIC for gauging individual learning is, in general, inefficient or wrong. We have seen that, in a teaching program that totaled 53 hours, the variability of TOEIC results defeated their usefulness in measuring learning gains because the SEM of TOEIC was in the range of expected individual gains.

4. Counseling learners on their progress. *Bad.* Because of the SEM of TOEIC, test-to-test differences will display very great variability. For example, differences may be negative, or they may be very large—and somewhat illusory in both cases. Indeed, the lower results that are frequently encountered in successive tests can have the unfortunate side effect of demotivating learners.

5. Guiding the courses of study of individual learners. *Bad.* TOEIC is not a diagnostic test, and it cannot pinpoint learners' strengths and weaknesses. It can be a rough guide for gauging a learner's overall level, if the administrator clearly understands the statistical variability of the results, but TOEIC cannot help the administrator determine *what* a specific learner needs to be taught.

No matter how loudly we proclaim that TOEIC should not be used for purposes for which it is not suited, proclaiming alone is not a long-term solution. Having no practical alternative, education directors will continue to use TOEIC because it is familiar and, if inefficient, at least a known evil. Using TOEIC may be better, for instance, than allowing language schools and internal education departments to devise their own tests, for that, to paraphrase Saegusa (1983), would be putting the fox in charge of the hen house. Nevertheless, company education directors and language schools should be warned that short-term TOEIC results cannot be substituted for more specific measures of learning achievement. Test users await a series of criterion-referenced tests complementary to the norm-referenced TOEIC. The new series of tests would consist of achievement tests that can be integrated with curriculum and can indicate mastery of or need for specific types of learning modules.

Because of these conclusions, company education directors who incorporate TOEIC into their testing programs should do so thoughtfully. They should understand that the long-term solution to many of their needs will be not TOEIC but a series of tests that are in tune with the specific goals and methods of their English education programs. TOEIC should be used in its area of strength, which is to find the approximate location of learners on the global bell curve of English proficiency.

Notes

[1] I gratefully acknowledge the assistance of JD Brown, Atsushi Kodera, and Rory Rosszell in helping me achieve what understanding I have of the use of TOEIC tests by Japanese companies.

[2] The approximation used for relating TOEIC standard scores to raw scores was:

Standard Score = (6.25 x Raw Score) - 225.

[3] Some experiences reported by Kitaoka (1992) suggest an even greater level of difficulty in the 700 - 800 range: about 3.6 teaching hours to raise a TOEIC score one point.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Brown, J. D. (1988). *Understanding research in second language learning*. Cambridge: Cambridge University Press.

Brown, J. D. (1995). *Testing in language programs*. Englewood Cliffs, NJ: Prentice-Hall.

ETS (Educational Testing Service). (1980). Score conversion tables, Form 3BIC, in *Test of English for International Communication* (p. 153). Princeton, NJ: Educational Testing Service.

ETS (Educational Testing Service). (1990). *Guide for TOEIC users*. Princeton, NJ: Educational Testing Service.

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). New York: McGraw-Hill.

Kitaoka, Y. (1992, July 30). How American firms in Japan should use TOEIC (Talk before the American Chamber of Commerce in Japan).

Saegusa, Y. (1983). TOEIC and in-company English training. *Cross Currents, 10*(1), 71-89.

Saegusa, Y. (1985). Prediction of English proficiency progress. *Musashino English and American Literature, Vol. 18*. Tokyo: Musashino Women's University.

TOEIC Newsletter (quarterly). Tokyo: Institute for International Business Communication.

TOEIC Steering Committee (1991). *Test of English for international communication (TOEIC): History & Status*. Princeton, NJ: Educational Testing Service.

Woodford, P. E. (1980). *The test of English for international communication (TOEIC)*. Paper presented at English-Speaking Union Conference, London, England.

Chapter 9

A Comparison of TOEFL and TOEIC

SUSAN GILFERT
AICHI PREFECTURAL UNIVERSITY

What are these two tests, the *Test of English as a Foreign Language* (TOEFL) and the *Test of English for International Communication* (TOEIC)? What do they test? Are they useful? How are they used? What do the scores mean? Are the scores comparable in any way? These and other questions will be answered in this paper. Knowledge of the background of each test will help users understand the test so a short history of each test will be provided, and a general structural comparison is made; then examples of question types will be given, subsection by subsection. A brief discussion will also be provided of how results of the exams are used or misused.

Comparative History of TOEFL & TOEIC

The TOEFL was generated by American academic needs in the early 1960s. American universities had many applications from students whose native language was not English, and the university administrators needed to know what to do with these students. ESL programs were ongoing everywhere, but without much cohesion. University administrators were familiar with the reputation and commercial value of the Educational Testing Service (ETS) tests because the Scholastic Achievement Tests (SATs) produced by ETS were being used successfully by university admissions offices. Universities requested ETS to create a general test of English "to evaluate the proficiency of people whose native language is not English" (ETS, 1993), specifically North American English. Today, it is used primarily in U.S. universities for admissions decisions, and sometimes misused for placement testing purposes in ESL settings. Most of the examinees are in their mid-teens to mid-twenties and are high school or university students.

The TOEIC came from the Japanese Ministry of International Trade and Industry requests in the middle 1970s. It is "designed to measure the English-language listening comprehension (LC) and reading (R) skills of individuals whose native language is not English. The TOEIC is used primarily by corporate clients, worldwide" (Wilson, 1989). Most examinees are in their mid-twenties to late forties and work for a corporation. From its beginning nearly 20 years ago, the use of TOEIC has spread from Japan throughout Asia, and it is becoming more frequently used throughout Europe and South America.

Structure Comparison

The TOEFL is a well-known multiple-choice instrument designed to measure an examinee's receptive English skills and is considered a reasonably good predictor of the examinee's productive language skills. The general register of the TOEFL is academic English. The TOEIC is a lesser-known multiple-choice instrument designed to measure a examinee's receptive English skills, and is increasingly becoming considered a reasonable predictor of these skills. The general register of the TOEIC is real-life, business-type English. The TOEFL is created, produced, and sold by Educational Testing Services in Princeton, New Jersey; the TOEIC was created by ETS, but is now entirely owned and operated by the Japanese TOEIC office in Tokyo.

The two tests are not radically different in structure (see Table 1). The topic treated in each test, however, is different. For example, in the reading comprehension subtests, the TOEFL uses passages which are purported to be found in first-year college textbooks; whereas the TOEIC tends to use business letters, short news items, and advertisements. However, the type of questions asked in the Reading Comprehension subsection (main idea, inference, attitude/tone, or application within the passage) are similar in the two tests. This same relationship (topic/type of question) exists in the Short Talks subsections. The Incomplete Sentence and Error Recognition subsections are also almost completely alike.

Listening Comprehension

Specific comparisons of listening sections. Section 1 of the TOEIC measures the ability of the examinee to recognize vocabulary in the context of a photo prompt (see Table 2). The TOEFL does not have any equivalent to this type of question. Examinees tend to feel that the photo prompt, providing visual context, is reassuring, even though both the question and possible answers are only spoken, not printed.

Section 2 of the TOEIC assesses the examinee's ability to listen to a question and choose the appropriate response (see Table 3). Some Japanese examinees have commented that this section seems to be mostly a structure test, listening for the grammatically

Table 1. *General comparison of tests*

TOEFL	TOEIC
3 major subtests; 5 subsections 150 questions	2 major subtests; 7 subsections 200 questions
scaled score ranges from 200 to 677	scaled score ranges from 10 to 990
examinees tend to be students (18-25 years old)	examinees tend to be corporate-level employees (25-50 years old)
results tend to determine schools to be attended and other academic matters	results tend to determine overseas postings and other business related matters
I. Listening Comprehension	I. Listening Comprehension
	A. One photograph, spoken sentences (20 Qs)
	B. Spoken utterances, spoken response (30 Qs)
A. Short conversation (25 Qs)	C. Short conversation (30 Qs)
B. Short talks (25 Qs)	D. Short talks (20 Qs)
II. Struct/Written Expression	II. Reading Comprehension
C. Incomplete sentences (15 Qs)	E. Incomplete sentences (40 Qs)
D. Error recognition (25 Qs)	F. Error recognition (20 Qs)
III. Reading Comprehension	G. Reading comprehension (40 Qs)
E. Reading comprehension (30 Qs)	

Table 2. *Question example comparison: listening comprehension*

TOEFL	TOEIC
no comparable section on TOEFL	A. One photograph, spoken sentences (Gilfert & Kim, 1995): Seen: A photo of two men talking across a table. An unused computer is in the background. Heard: (A) The two men are computing. (B) The computer is having a meeting with the men. (C) The two men are talking. (D) One man is buying a computer. (C) is the correct answer since it is closest in meaning tto what is shown in the photo.

correct response. Most examinees feel that this part of the TOEIC is the most difficult part of the listening component since both the prompt and the possible answers are only spoken, not printed.

Section 3 of the TOEIC is most similar to Section 1 of the TOEFL listening subtest (see Table 4). The TOEFL "short conversation" subsection of the tape follows an A: B: pattern, then a Narrator asks a question about the conversation. Four printed possible answers appear in the test booklet. The TOEIC "short conversation" tends to follow an A: B: A: pattern where the question and four possible answers are printed in the test booklet. Examinees tend to feel that the TOEIC is easier to understand in this section because both the question and possible answers are printed in the test book, which provides examinees more context into which to fit the conversation. In both tests, there is only one question per conversation.

Section 4 of the TOEIC is most similar to Section 2 of the TOEFL listening subtest, but the TOEIC talks tend to be shorter, and there tend to be fewer questions for each talk (see

Table 3. *Question example comparison: listening comprehension*

TOEFL	TOEIC
Not comparable; look at next section	B. Spoken utterances, spoken response (Gilfert & Kim, 1995): Heard: Hello, John. Heard: (A) Hi, John. How are you? (B) Who's John? (C) Good-bye, see you later. The correct response is (A), since it is the most likely response to this greeting.

Table 4. *Question example comparison: listening comprehension*

TOEFL	TOEIC
A. Short conversations (adapted from ETS, 1993, p. 26):	C. Short Conversation, Four printed answers (Gilfert Kim, 1995):
Heard: Man: Do you mind if I turn off the television? Woman: Well, I'm watching this show. Narrator: What does the woman imply?	A: May I help you? B: Yes, do you have this shirt in size 12? A: Certainly. I'll get one for you.
Read: (A) The man should show his watch to the woman. (B) The man should leave the television on. (C) The show will be over soon. (D) The woman will show the television to the man.	Read: Where is this conversation most likely taking place? (A) in a hotel (B) in a department store (C) in a post office (D) in an airport
The correct response is (B), since the woman implies that she is using the television.	The correct response is (B), since the conversation appearsto be happening between a sales clerk and a customer.

Table 5). The TOEFL "longer conversations" subtest normally have up to 1.5 minutes' worth of spoken language and ask 4 to 6 questions, whereas the TOEIC "short talks" subtest tends to have shorter talks (1-1.5 minutes) and ask 3-5 questions per talk. The content of the TOEFL talks are typically either long conversations (4-10 exchanges) between two people talking about school-related matters, or single speakers giving an introduction to a class or to a school club activity. The content of the TOEIC is typically made up of extended conversations (5 or 6 extended exchanges) between two people talking about office matters, or single speakers giving a news report or other information. Examinees tend to feel that the TOEIC material is less difficult than the TOEFL since both the amount of spoken language and the number of questions are reduced.

General comparison of listening subtests. Comparing the listening subtests in a general sense, there are 50 questions on tape in the TOEFL and 100 questions on tape in the TOEIC. The listening subtest of the TOEFL requires about 30-40 minutes to take; the Listening subtest of the TOEIC takes about 40-50 minutes. Vocabulary is tested throughout both tests; it is no longer a separate section. The general register of the TOEFL listening subtest is academic with the topics of conversation tending to be something about school or everyday life. The general register of the TOEIC listening subtest is "business" with more idioms being spoken than on the TOEFL and fewer polysyllabic words.

Grammar

The next two sections of the TOEIC and the TOEFL are identical in structure; the only difference is the register. These two sections are: Incomplete Sentences and Error Recognition. They both assess the examinee's knowledge of English structure, or grammar. The

Table 5. *Question example comparison: listening comprehension*

TOEFL	TOEIC
B. Longer Conversations (Gilfert& Kim, 1990):	D. Short Talks (Gilfert & Kim, 1995):
Heard: Questions 3, 4, 5, and 6 are based on the following conversation: Woman: What do you think? Am I OK? Man: Exhale slowly, please. Well, there is some congestion. I want to do some tests. Woman: How soon will I get the results? Man: Oh, you'll have the results before you leave the office, and here is some medicine that I believe will help you.	Heard: Sunshine is forecast for today after two damp days. Westerly winds will freshen by afternoon and chilly air will be transported across the metropolitan area. Clouds will overtake clear skies by morning.
Question 3: What is the probable relationship between these two speakers?	Read: 1. How was the weather earlier this week? (A) Sunny C) Damp (B) Cool and dry (D) Chilly
Question 4: When will the woman receive the results of the tests?	2. What kind of weather is expected tomorrow? (A) Cool and cloudy (C) Damp and windy (B) Sunny and dry (D) Cold and sunny
Question 5: What does the man feel will help the woman?	For question 1, (C) is the best answer. The announcer notes that the last two days have been damp.
Question 6: What is the woman's problem?	For question 2, (A) is the best answer.
Read: 3. (A) dentist-patient (B) doctor-patient (C) teacher-student (D) pharmacist-customer	
4. (A) in a few days (B) before leaving the office (C) very slowly (D) soon enough	
5. (A) some medicine (B) some tests (C) exhaling slowly (D) filling her lungs with air	
6. (A) She does not have enough air in her lungs. (B) She's exhaling too slowly. (C) She didn't do well in her tests. (D) She has a little congestion.	
For question 3, (B) is the best answer. For question 4, (A) is the best answer. For question 5, (A) is the best answer. For question 6, (D) is the best answer.	

Table 6. *Question example comparison: grammar*

TOEFL	TOEIC
C. Incomplete Sentences (adapted from ETS, 1993, p 27):	E. Incomplete Sentences (Gilfert & Kim, 1995):
Read: The fact ____ traveller's checks can usually be easily changed to cash makes them a popular way to carry currency.	Read: ____ girl over there is my sister.
(A) of (C) is that (B) that (D) which is	(A) This (C) Those (B) These (D) That
(B) is the most grammatically correct answer.	(D) is the answer that is grammatically correct here.

TOEFL subtest is supposed to "measure ability to recognize language that is appropriate for standard written English" (ETS, 1993), and the TOEIC subtest measures much the same. The examples in Table 6 both test demonstrative pronoun usage. Either example would work for TOEFL or TOEIC. As noted before, this section on each test is practically identical. Register may differ slightly, but not significantly.

The examples in Table 7 both test word order. Again, either example would work for TOEFL or TOEIC. As noted before, this subtest on each test is practically identical. In comparing the grammar subtests, there are 40 questions in the Grammar subtest of the TOEFL. There are 60 questions in the Grammar subtest of the TOEIC. For timing, examinees should allow about 25 seconds for each question in this subtest, taking about 15 or 20 minutes for the TOEFL and about 25 minutes for the TOEIC. However, the Grammar and Reading Comprehension sections are timed together. If an examinee can quickly (and accurately) go through the Grammar section, then more time is left for the Reading Comprehension questions.

Reading Comprehension

The examples in Table 8 clearly show the difference in register between TOEFL and TOEIC Reading questions. The TOEFL reading example is likely adapted from some textbook, and the TOEIC example is a news

Table 7. *Question Example Comparison: Reading Comprehension*

TOEFL	TOEIC
D. Error Recognition (adapted from ETS, 1993, p. 28):	F. Error Recognition (Gilfert & Kim, 1995):
Read: Good puzzles provide an A B excellent way to explore the C area of thought abstract. D	Read: In today's class middle, A both parents have to work in B C order to pay all their bills. D
(D) is the error here; the words are in the wrong order.	(A) is an error in word order, making it the correct answer.

Table 8. *Question example comparison: reading comprehension*

TOEFL	TOEIC
Questions 40-41 refer to the following passage (Gilfert & Kim, 1990, p. 93): In 1920, after some thirty-nine years of problems with disease, high costs, and politics, the Panama Canal was officially opened opened, finally linking the Atlantic and Pacific Oceans by allowing ships to pass through the fifty-mile canal zone instead of traveling some seven thousand miles around Cape Horn. it takes a ship approximately eight hours to complete the trip through the canal and costs an canal and costs an average of fifteen thousand dollars, one-enth of what it would cost an average ship to round the Horn. More than fifteen thousand ships pass through its locks each year. The French initiated the project but sold their rights to the United States. The latter will control it until the end of the twentieth century when Panama takes over its duties. 40. According to the passage, who currently controls the Panama Canal? (A) France (C) Panama (B) United States (D) Canal Zone 41. In approximately what year will a different government take control of the Panama Canal? (A) 2000 (C) 3001 (B) 2100 (D) 2999 40. (B) is the correct answer. 41. (A) is the correct answer.	Questions 1-2 refer to the following report following report (Gilfert & Kim, 1995): SAN FRANCISCO (NNN)–Rains, accompanied by high winds, closed a number of schools on Monday, but the storm was welcomed because it brought some relief from a fifth straight year of drought. 1. What has the weather been like in California for the last five years? (A) wet (C) dry (B) windy (D) mild 2. Why would people want rain? A) California people enjoy walking in the rain. (B) California roads need relief relief from the sun. (C) California school children want to study rain. (D) There has been no rain for five years. 1. (C) is the correct answer. 2. (D) is the correct answer.

report. However, the types of reading comprehension questions are mostly the same on each test: main idea, details, inference, and/or author's attitude.

The TOEFL reading comprehension subtest tends to have 3-5 questions per reading passage, and the TOEIC tends to have 3-4 questions per passage. In comparing the reading subtests, there are 30 questions in the Reading subtest of the TOEFL. There are 100 questions in the second section of the TOEIC: 60 questions in the Grammar subtest and 40 questions in the Reading subtest. The TOEIC Grammar/Reading sections are timed together, and the examinee is free to switch back and forth between subtests. The Grammar and Reading subtests of the TOEFL require 80 minutes to take; the Grammar/Reading subtest of the TOEIC takes 75 minutes.

Purposes of the Tests

Both the TOEIC and the TOEFL are useful tests, but each is used for different purposes. The stated purposes of both tests are to provide a general measure of the examinees' English ability. However, some institutions misuse the tests for purposes which should not be measured on these tests. For example, the TOEFL is used correctly as a measure of an examinee's overall ability to use academic English or to determine whether a examinee may enter full-time study at a university. Sometimes, however, TOEFL results are incorrectly used for placement purposes in EFL settings or as determiners for employment or placement within a company. Many Japanese universities and colleges regularly offer TOEFL-preparation courses to their students. These students may be thinking about going to graduate schools overseas, or simply need to get a high TOEFL score in order to get a job.

The TOEIC is correctly used to assess an examinees' overall English proficiency in a business context. TOEIC scores are increasingly being required by corporate employers of either entering employees or employees who are being considered for promotion and/or overseas assignments. Employers often use TOEIC scores as a screening device, hiring only those who meet a certain pre-determined TOEIC score. As a result of this practice, Japanese colleges, universities, and tertiary-level vocational schools are now offering TOEIC-preparation courses in greater numbers than five years ago. TOEIC-preparation courses have already been offered by language schools throughout Japan for many years now. Some corporate employers use the TOEIC incorrectly, by requiring their domestic employees (who do not use English on a regular basis) to obtain a certain score for promotion or raises.

Score Comparison

Some researchers and students of testing believe that the TOEFL shows the difference between intermediate-to-advanced levels in English proficiency very well, but does not discriminate well on scores lower than about 450 (generally considered intermediate-level). For most U.S. academic purposes, a TOEFL score of 500 is acceptable for a student to begin part-time studies, along with some ESL courses. A TOEFL score of 550 is generally considered acceptable for a non-native English speaker to do undergraduate studies full-time, and a TOEFL score of 600 is generally accepted for full-time graduate studies.

The TOEIC, on the other hand, is believed to show the differences between low-beginner to high-intermediate levels very well. A TOEIC score of 450 is frequently considered acceptable for hiring practices, with the understanding that the employee will continue English studies. A TOEIC score of 600 is frequently considered the minimum acceptable for working overseas. Domestic-based engineers who have a TOEIC score of 500 are considered reasonably proficient in English. If the same engineer is being considered for a posting overseas, he or she must usually try for a TOEIC score of about 625. A domestic-based desk-worker with a TOEIC score of 600 is considered reasonably proficient in English. For the same desk-worker to go overseas, she or he must usually have a TOEIC score of 685.

What do the Scores Mean?

The TOEIC office in Tokyo, Japan, has published a comparison between the Oral Proficiency Index (the OPI is used by the U.S. Foreign Service), TOEFL scores, TOEIC scores, other tests, and the Japanese *Eiken*. All these tests assess English reading, listening, and grammar proficiency. The OPI and *Eiken* series further test speaking ability. The Oral Proficiency Index is considered one of the best tests since it provides a means of testing the examinees' productive language skills, as well as their receptive language skills. However, due to time and cost considerations, the OPI is an impractical test to administer for large numbers of people. The Educational Testing Service also administers the *Test of Spoken English* (TSE) and the *Test of Written English* (TWE). Of the 12 official (International) TOEFL tests administered every year, five include a TWE in addition to the regular TOEFL. The TSE has its own testing schedule, since it requires making an audio tape. Many schools do not require these additional tests for much of their admissions procedures, although non-native English speaking graduate students who wish to become Graduate Teaching Assistants are increasingly required to pass the TSE in order to get an assistantship. Both TOEFL and TOEIC test receptive skills (listening and reading) rather than productive skills (speaking and writing). It is possible for students to score very high on the TOEFL, but not be able to use oral or written English in context. Many examinees become expert in taking language tests, but do not learn how to use the language. Therefore, the author maintains that TOEFL and TOEIC tests operate in an "artificial reality."

The tests, when used alone, are valid and reliable in themselves, but not in a larger sense. Examinees who score well on these tests may have self-confidence in the language classroom, but using their language skills in the real world may be quite a different thing. In theory, examinees with a TOEFL score of 550 should be able to use their receptive English skills better than examinees with a TOEFL score of 500. But the examinee's English will also be used productively in the actual academic world. Hence, their English listening abilities have to be good enough to permit them to write notes in class, and their writing abilities must typically be equal to English native speakers' when creating papers or reports. Similarly, the language used in the TOEIC is intended to be used in receptive language contexts. An examinee with a score of 650 would be expected to operate in an English-speaking business context better than an examinee with a score of 600. In the real world, examinees will be reading and generating faxes and reports, listening to and making presentations, and using the telephone. Examinees who excel in taking paper tests, yet are unable to use their language productively, will be at a loss in the real world.

Are the Scores Comparable?

Finally, the question comes to this: What is the difference between the TOEFL and the TOEIC? Can they be compared at all? The scoring system is different and the number of questions is different, as is the amount of time needed to take each test. The register is also different. The reasons for taking each test (the examinees' motivation) can be different (except perhaps in the area of securing employment), and the ways of using the results of the tests are different. The vocabulary in the two tests has areas of similarity, but there are some noticeable differences. Some examinees feel that the TOEIC is easier than the TOEFL. Some students of testing consider that the TOEFL is a more accurate discriminator for higher-level examinees, and the TOEIC is a more accurate lower-level discriminator. The tests were both created by Educational Testing Service, and test American English. ETS has calculated a number of reliability and validity checks on both tests, so they are both considered accurate and useful when used within the guidelines published by ETS. The grammar subtests of both tests are quite similar and the types of questions asked in the Reading Comprehension subtest (main idea, details, inference, and/or author's attitude) are similar. In short, with proper understanding of the TOEFL and the TOEIC, they can be useful, but they must be used properly, with full knowledge of their limitations.

Notes

[1] The Japanese *Eiken* is a six-part series of non-standardized tests produced by the Society for Testing English Proficiency (STEP), published in Tokyo, Japan. The STEP test levels are level 4 (low beginning), 3 (high beginning), pre-2 (low intermediate), 2 (high intermediate), pre-1 (low advanced) and 1 (high advanced). STEP is offered by preregistration at a relatively low cost (¥2,500 -¥3,000 per person per test) twice a year in Japan. There is no limit on how often a person can take the test. In contrast, the International TOEFL is available, by preregistration only, 12 times a year, anywhere in the world. The cost of taking a TOEFL is U.S. $42 on a

Friday administration or $35 on a Saturday administration, payable in U.S. (or Canadian) funds only. Five times a year, the TWE is part of the TOEFL at no extra cost. The International TOEFL is the official TOEFL; the scores are sent at the examinee's request to examinee-selected schools. Five of the International TOEFL administrations are disclosed. Examinees can choose to give a self-addressed envelope and postage for 43 grams from the U.S. to the test administrators. The examinee's test booklet will be mailed a week or so later to the examinee. There are other versions of the TOEFL, called Institutional TOEFL. Institutions may choose to purchase and offer the TOEFL to their students or employees at any time they wish, as long as that date does not conflict with an International TOEFL. These scores are used within the institution which offers the TOEFL; the scores do not leave the institution for use in applying to schools in the U.S. or Canada. The TSE is offered at other times, for U.S. $80 (for TSE-A) or $110 (for TSE-P). The TOEIC is offered, by preregistration only, six times a year for ¥6,500, payable in Japanese yen or the equivalent in local funds. This information may change. Please check the latest bulletin of information for the latest prices, test dates, and availability. Bulletins are available free of charge at many bookstores and university or college campuses.

References

Gilfert, S., & Kim, V. (1990). *TOEFL Strategies.* Nagoya, Japan: Kawai Juku Press.

Gilfert, S., & Kim, V. (1995). *TOEIC Strategies.* Tokyo: Macmillan Japan K. K.

ETS. (1993). *TOEFL 1993-4 bulletin of information.* Princeton, NJ: Educational Testing Service.

Wilson, K. M. (1989). *TOEIC research report #1.* Princeton, NJ: Educational Testing Service.

Chapter 10

English Language Entrance Examinations at Japanese Universities: 1993 and 1994

JAMES DEAN BROWN
UNIVERSITY OF HAWAII AT MANOA

SAYOKO OKADA YAMASHITA
INTERNATIONAL CHRISTIAN UNIVERSITY

As we have pointed out in more detail in Brown and Yamashita (1995), prestigious secondary schools and universities in Japan administer their own entrance examinations on a yearly basis to students who would like to be admitted to their student bodies. To prepare for these examinations students spend months and years working industriously in school, at home, and in jukus in a kind of language testing hysteria (Brown, 1993). The whole process is known as shiken jigoku, or examination hell. Examination hell is not a new phenomenon. Amano (1990, p. xx) says that it existed in its present form in the 1920s with roots back to the Meiji Restoration. Because most Japanese believe that the success of their children hinges on passing these examinations ". . . families devote a surprising proportion of their resources toward assisting their children in exam preparation, and children devote long hours day after day to study" (Cummings, 1980, p. 206).

Many Japanese are not 100 percent happy with the current examination system. Frost (1991) gives an overview of ways that Japanese criticize, need, admire, and tolerate the entrance examination system. Tsukada (1991, p. 178) explains ways in which these examinations have "undesirable effects on curriculum, on foreign language instruction, on family life, and on children's emotional, physical, and intellectual development." As Frost (1991, p. 303) noted, however, "separate university achievement tests simply have too long a history and meet too many needs . . . to disappear simply because a new generation is beginning to be truly worried about them." Since this examination system so dramatically affects English teaching in Japan, we feel that EFL teachers should know as much as possible about it.

To that end, we published Brown & Yamashita (1995), a study of the English language entrance examinations at 21 Japanese universities in 1993. The present study expands on that project by describing the examinations produced by the same universities in 1994 and comparing the 1993 and 1994 examinations.

Definitions

Before looking at those examinations in detail, we will briefly define some key terms (fuller explanations are given in Brown & Yamashita, 1995):

Test item. The smallest distinctive unit on a test that yields separate information, that is, the smallest part that could be scored separately (after Brown, 1995). For instance, a single multiple-choice item is a test item, but so is an essay, or a translation task.

Discrete-point vs. integrative. A discrete-point item is one that is designed to test a single well-defined language point; for example, true-false, multiple-choice, matching, and sometimes fill-in items are usually discrete-point in nature. Integrative test items are harder to define and identify because they are situated in a language context and because they may interact with each other and other elements of the language context in relatively complex ways; for instance, dictations, cloze tests, essay writing tasks, and interviews, are made up of integrative items.

Receptive vs. productive. Receptive item types are those in which the students receive language in the sense of reading it or hearing it, but are not required to produce language (i.e., are not required to speak or write anything); generally, true-false, multiple-choice, and matching items are receptive. Productive items require the students to actually produce language in one form or another; fill-in, short answer, and task-oriented (e.g., compositions, interviews, role plays, etc.) are productive.

Translation items. Translation items require the students to translate a phrase, sentence, paragraph, etc. from their first language (L1) into their second language (L2), or vice versa; for instance, a student might be required to write out in Japanese the translation of a sentence in English. We argue in Brown and Yamashita (1995) that translation may be too difficult, demanding, and specialized a task to require of students who have only studied a language through junior and senior high school, and that putting translation tasks on a test signals the students that it is a legitimate strategy to use in normal communication.

Purpose

The purpose of this study is to further investigate the Japanese university English entrance examinations by examining some of those produced in 1994. To that end, we have set ourselves two goals. Our first goal is to describe the state of affairs in 1994 at some of the major universities in Japan in order to better understand how these examinations vary from university to university. Our second goal is to compare the 1994 examinations with those produced in 1993 to determine how the examinations vary from year to year.

To those ends, the following more formal research questions were posed:

1. How difficult are the various reading passages used in the 1994 university English language entrance examinations?
2. Are there differences in the levels of reading passage difficulty in private and public university examinations in 1993 and 1994?
3. What types of items are used on the 1994 English language entrance examinations, how varied are they, and how does test length vary?
4. Are there differences in the types of items and test lengths found in private and public university examinations in 1993 and 1994?
5. What skills are measured on the 1993 and 1994 English language entrance examinations?

It is hoped that answering these questions will help English teachers in Japan prepare their students for such tests and encourage responsible testing practices at the various universities that do this kind of testing.

Method

Materials

Two books served as our primary sources: Kôkô-Eigo Kenkyû, 1993a; Kôkô-Eigo Kenkyû, 1993b. Both were readily available at commercial bookstores. Each book contains a number of English language entrance examinations along with hints in Japanese on studying for these examinations. Kôkô-Eigo Kenkyû, 1994a contained examinations from 67 *private universities*. The same ten universities used in Brown and Yamashita (1995) were selected for this study: Aoyama University, Doshisha University, Keio University, Kansai Gaidai (Foreign Languages) University (KANGAI), Kansai University, Kyoto University of Foreign Studies (KYOTOUFS), Rikkyo University, Sophia University, Tsuda University, and Waseda University. Kôkô-Eigo Kenkyû, 1994b contained examinations from 56 *public university* examinations, 41 of which were from national universities, 15 from municipal universities, and one was the nationwide "center" examination. The same ten public universities used in our previous study were selected here. Eight were national universities: Hitotsubashi University (HITOTSU), Hokkaido University, Kyoto University, Kyushu University, Nagoya University, Osaka University, Tokyo University, and Tokyo University of Foreign Studies (TYOUFS). And, two were municipal: Tokyo Municipal University (TORITSU) and Yokohama City University.

The *Daigaku Nyuushi Sentaa* (or so-called *center* examination), which is administered nationwide, was also included in both studies for a total of 21 examinations (see Brown & Yamashita 1995, p. 12, for more information on the *center* examination).

We originally selected ten each of the most prestigious private and public universities with a view to also getting a good geographical distribution. The *center* examination was included because it is administered nationwide.

Analyses

We began by labeling each item for its type, its purpose, the number of options students had, the language(s) involved (Japanese or English), the task required, and any other salient features.

Data entry took several forms in this project, but all of the computer programs used were for IBM (MS-DOS) computers as follows: 1) Each item was coded for item type and recorded in the *QuattroPro*™ spreadsheet program (Borland, 1991); 2) all reading passages were typed into a word processing program using the *WordPerfect*™ (WordPerfect, 1988) computer program;[1] 3) the reading passages in the entrance examinations were analyzed using the *RightWriter*™ computer program (Que Software, 1990) for such features as number of words, number of unique words, percent of unique words, number of sentences, syllables per word, words per sentence, as well as the Flesch, Flesch-Kincaid, and Fog readability indexes. The number of words, percent of unique words, number of sentences, syllables per word, and words per sentence are self-explanatory, but some of the other characteristics may not be so clear. The number of unique words can also be described as the number of different words (i.e., no single word counts more than once). The Flesch, Flesch-Kincaid, and Fog readability indexes are all ways of estimating the reading-level difficulty of a passage. The Flesch scale ranges from 0 to 100 with a higher number indicating an easier passage. The Flesch-Kincaid and Fog readability indexes are both meant to show the grade level of students for which the reading passages should be appropriate. (For more on these readability indexes, see Brown & Yamashita, 1995; Que Software, 1990, pages 7.5 to 7.6; and Flesch, 1974).

Only descriptive statistics were used in this project to make comparisons between universities, types of universities, and the 1993 and 1994 results. These statistics consisted primarily of averages and percents. Because of the descriptive nature of this study, no inferential statistics were used.

Results

The results of this study address the original research questions posed earlier. As such, the questions themselves will be used as headings to help organize the discussion.

1. *How difficult are the various reading passages used in the 1994 university English language entrance examinations?*

Since all of the examinations in this study used at least some reading passages as the basis for items on the test, we began by examining those passages. Tables 1A and 1B present the statistics for the reading passages on the examinations at the private and public universities, respectively. Notice that the universities

TABLE 1A: READING PASSAGE STATISTICS FOR PRIVATE UNIVERSITIES, 1994 (AVERAGES)

STATISTIC	PRIVATE UNIVERSITIES AOYAMA	DOSHISHA	KEIO	KANGAI	KANSAI	KYOTOUFS	RIKKYO	SOPHIA	TSUDA	WASEDA
NO. OF PASSAGES	2	2	1	2	2	2	2	7	3	4
WORDS	445.50	622.00	914.00	423.50	778.00	471.50	507.00	410.57	332.67	565.75
UNIQUE WORDS	228.00	303.50	435.00	221.00	304.50	226.50	258.50	217.71	157.00	296.50
UNIQUE WORDS (%)	51.18	48.79	47.59	52.18	39.14	48.04	50.99	53.03	47.19	52.41
SYLLABLES/WORD	1.42	1.64	1.62	1.57	1.37	1.52	1.39	1.50	1.47	1.61
SENTENCES	27.00	29.00	36.00	23.50	55.50	29.50	22.50	19.71	16.67	28.50
WORDS/SENTENCE	17.75	19.91	22.29	19.84	14.20	15.62	21.81	20.41	23.86	19.44
FLESCH	69.09	47.90	46.85	53.96	76.93	62.52	67.11	59.29	58.52	51.39
FLESCH-KINCAID	8.04	11.53	12.26	10.66	6.06	8.42	9.32	10.06	11.03	10.92
FOG	10.58	14.12	14.56	12.42	8.64	9.98	11.82	11.84	13.75	12.85

TABLE 1B: READING PASSAGE STATISTICS FOR PUBLIC UNIVERSITIES, 1994 (AVERAGES)

STATISTIC	PUBLIC UNIVERSITIES HITOTSU	HOKKAIDO	KYOTO	KYUSHU	NAGOYA	OSAKA	TOKYO	TORITSU	TYOUFS	YOKOHAMA
NO. OF PASSAGES	2	4	2	3	2	3	3	2	7	4
WORDS	598.50	335.75	461.50	488.33	466.50	297.00	410.00	513.50	342.43	262.75
UNIQUE WORDS	317.00	177.00	234.50	258.33	267.00	184.33	210.67	247.00	176.14	156.25
UNIQUE WORDS (%)	52.13	52.72	50.81	52.90	57.23	62.07	51.38	48.10	51.44	59.47
SYLLABLES/WORD	1.45	1.43	1.45	1.47	1.68	1.57	1.42	1.48	1.47	1.59
SENTENCES	36.00	20.25	40.00	33.00	19.00	14.33	27.00	32.00	17.43	12.75
WORDS/SENTENCE	16.56	16.03	13.43	15.41	23.96	20.75	16.01	15.97	19.26	21.83
FLESCH	67.49	69.36	70.51	67.12	40.21	52.58	70.33	65.30	63.16	50.28
FLESCH-KINCAID	7.96	7.57	6.76	7.72	13.61	11.08	7.42	8.12	9.24	11.67
FOG	9.93	9.29	9.09	9.95	15.28	13.59	9.61	10.83	11.53	13.65

are labeled across the top of the columns, while the statistics are labeled for the rows.

Tables 1A and B indicate that all of the private and public universities used at least one passage and that one of the private universities, SOPHIA, and one of the public universities, TYOUFS, used as many as seven passages. The average lengths of the passages can be considered at the same time by examining the number of words. While SOPHIA and TYOUFS had numerous passages, they were shorter than the passages in most of the other universities at 410.57 and 342.43 words per passage on average, respectively. In contrast, one of the private universities, KEIO, used only one passage, but used relatively a relatively long one at 914 words.

The number of unique words represents the number of different words used in the passages because each word is counted only once in this statistic. The percent of unique words was calculated by dividing the number of unique words by the total number of words. The percent of unique words can be considered an indicator of the relative amount of variety in the vocabulary of the passages because it shows the proportion of words which were unique and is therefore easily comparable across passages, universities, and university categories. The number of syllables per word gives additional information about vocabulary in that it is a rough indicator of the difficulty of the words (based on the idea that longer words are usually more difficult than shorter ones).

The number of sentences per passage is relatively straightforward to interpret. More interesting perhaps is the notion of sentence length, which can be gauged by looking at the sentence length statistic, which describes the average length of the sentences. Average sentence length is often considered a rough indication of the syntactic complexity of a passage.

The Flesch readability index indicates that the passages in the various universities ranged in overall reading level from "fairly easy" (76.93) at KANSAI to "difficult" (40.21) at NAGOYA. The Flesch-Kincaid readability index indicates that the passages would be appropriate for native speakers ranging generally from early sixth grade (about 11 years old) at KANSAI to late thirteenth grade level (third year of university, or about 21 years old) at NAGOYA (13.61). The Fog index generally appears to agree with the Flesch-Kincaid one, but is consistently about two grades higher.

TABLE 2: READING PASSAGE STATISTICS SUMMARIZED BY UNIVERSITY TYPE (AVERAGES)

	1993 EXANS				1994 EXANS			
STATISTIC	PRIVATE	PUBLIC	CENTER	TOTAL	PRIVATE	PUBLIC	CENTER	TOTAL
NO. UNIVERSITIES	10	10	1	21	10	10	1	21
NO. OF PASSAGES	2.30	3.40	3	2.86	2.70	3.20	3.00	2.95
WORDS	540.15	378.33	178.33	445.87	547.05	417.63	368.00	476.89
UNIQUE WORDS	272.58	196.95	100.67	228.38	264.82	222.32	189.67	241.01
UNIQUE WORDS (%)	50.74	52.94	53.88	51.84	49.05	53.83	51.54	51.44
SYLLABLES/WORD	1.51	1.52	1.41	1.51	1.51	1.50	1.49	1.50
SENTENCES	30.09	21.18	9.33	24.86	28.79	25.18	24.67	26.87
WORDS/SENTENCE	19.03	20.18	17.01	19.48	19.51	17.92	18.77	18.72
FLESCH	60.40	58.29	70.35	59.87	59.35	61.63	61.91	60.56
FLESCH-KINCAID	9.38	10.06	7.67	9.62	9.83	9.11	9.29	9.46
FOG	11.18	12.19	10.17	11.61	12.05	11.28	10.83	11.63

2. *Are there differences in the levels of reading passage difficulty in private and public university examinations in 1993 and 1994?*

The summary statistics shown in Table 2 reveal some interesting overall (average) differences in 1993 and 1994 between the private and public universities and the *center* examination. Notice that the overall averages are also given for all 21 universities considered together (TOTAL).

In both years, the public universities appear to have more reading passages, but shorter ones than the private universities, while the *center* examination is somewhere in between in terms of the number of passages, but with considerably shorter passages. Notice that on average the passages in every category are longer in 1994 than in 1993. So, it appears that there is a tendency for passages on the examinations to get longer.

In terms of percentage of unique words, syllables per word, and words per sentence, there are no particularly interesting differences between private, public, and *center* examinations, or between the 1993 and 1994 examinations. The average readability levels of the passages involved in these examinations vary, however, by type of university. The Flesch, Flesch-Kincaid, and Fog readability scales all indicate that the passages became slightly easier overall between 1993 and 1994. Examining the results in more detail reveals that the passages on the private university and *center* examinations became slightly more difficult on average between 1993 and 1994, while the passages in the public university examinations became slightly easier, on average.

3. *What types of items are used on the 1994 English language entrance examinations, how varied are they, and how does test length vary?*

Item types. Tables 3A and 3B summarize the different types of items that we found on the 1994 entrance examinations. Table 3A does so for the private universities, and Table 3B for the public ones. Notice that, as in the previous tables, the university names are provided across the top as column labels. Then, down the left-hand side, other labels indicate that the upper portion of each table contains frequencies while the lower part gives the equivalent percents. The frequencies are provided so that the actual tallies (or frequencies) can be examined by readers, but the percents are also given so that readers can make easier comparisons between universities (without the confusion caused by differing test lengths).

Notice further that, within the frequencies and percents, item types are categorized by skill area with reading and writing collapsed together to represent written tests, translation treated as a separate skill (as discussed above in the *Definitions* section), and listening handled as yet a third skill. Within each skill area, we found a variety of different types of items. Within the reading/writing skills, we found multiple-choice, true-false, rephrasing or reordering, fill-in the blank, and short answer or essay items. Within translation, we found both English to Japanese translation and Japanese to English. Among the listening items, we found true-false, multiple-choice, fill-in, dictation, and short-answer items.

Item variety. Tables 3A and 3B provide a great deal of information. Naturally, we cannot hope to interpret all of that information here in prose. Some salient facts do, however, stand out in these two tables.

For example, notice that different universities use different combinations of item types. Some universities, like those labeled KANGAI and SOPHIA, place great emphasis on multiple-choice items, while other universities like KYOTO place heavy emphasis on translation. Notice also that other universities, like AOYAMA, TOKYO, and TYOUFS, prefer to put a fairly heavy emphasize on listening items, while the other universities do not. The reader should continue to explore Tables 3A and 3B to discover how very different the examinations are from university to university. The one thing that seems clear, at this point, is that the nature of the item types on the various university entrance examinations varies tremendously. Thus the types of items that a student will face depends to a great degree on which examination the student chooses to take.

As in all classifying and categorizing, however, this process of analyzing the item types oversimplifies reality. For example, within each of our seemingly simple categories, there were additional differences. Consider just the first category of reading/writing multiple-choice items. There was tremendous variation in this seemingly straightforward item type. For instance, in terms of the number of options supplied in the items, we found them ranging from three to five. Some of the multiple-choice items had the options in English, some in Japanese. Some of the items were based on the reading passages (in a wide variety of different ways) and some stood alone. Some items posed a straightforward question, others required the students to select the option that successfully filled a blank. Still others required the students to select the option that matched a word or phrase, and finally, some provided context for blanks in the form of multiple-choice cloze. In addition some of the items were

TABLE 3A: THE VARIETY OF ITEM TYPES ON PRIVATE UNIVERSITY EXAMINATIONS, 1994

SKILL:	PRIVATE UNIVERSITIES									
ITEM TYPE	AOYAMA	DOSHISHA	KEIO	KANGAI	KANSAI	KYOTOUFS	RIKKYO	SOPHIA	TSUDA	WASEDA
FREQUENCIES										
READING/WRITING:										
MULTIPLE-CHOICE	10	34	0	79	43	43	23	80	8	17
TRUE-FALSE *	0	0	0	0	0	0	0	0	5	0
REPHRASE/REORDER	0	3	0	0	0	0	0	0	0	0
FILL-IN	0	6	0	0	0	0	14	0	31	14
SHORT-ANSWER/ESSAY	1	0	4	0	3	1	1	0	0	1
TRANSLATION:										
TRANSLATE (E->J)	2	1	3	0	2	1	1	0	2	0
TRANSLATE (J->E)	2	1	1	0	0	3	0	0	2	0
LISTENING:										
TRUE-FALSE	0	0	0	0	0	0	0	0	0	0
MULTIPLE-CHOICE	10	0	0	0	0	0	0	0	0	0
FILL-IN	0	0	0	0	0	0	0	0	0	0
DICTATION	0	0	0	0	0	0	0	0	1	0
SHORT-ANSWER *	0	0	0	0	0	0	0	0	0	0
TOTAL NO. OF ITEMS	25	45	8	79	48	48	39	80	49	32
TIME ALLOWED	100	100	120	90	90	80	95	90	100	90
PERCENTAGES										
READING/WRITING:										
MULTIPLE-CHOICE	40.00	75.56	0.00	100.00	89.58	89.58	58.97	100.00	16.33	53.13
TRUE-FALSE *	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.20	0.00
REPHRASE/REORDER	0.00	6.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FILL-IN	0.00	13.33	0.00	0.00	0.00	0.00	35.90	0.00	63.27	43.75
SHORT-ANSWER/ESSAY	4.00	0.00	50.00	0.00	6.25	2.08	2.56	0.00	0.00	3.13
TRANSLATION:										
TRANSLATE (E->J)	8.00	2.22	37.50	0.00	4.17	2.08	2.56	0.00	4.08	0.00
TRANSLATE (J->E)	8.00	2.22	12.50	0.00	0.00	6.25	0.00	0.00	4.08	0.00
LISTENING:										
TRUE-FALSE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MULTIPLE-CHOICE	40.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FILL-IN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DICTATION	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.04	0.00
SHORT-ANSWER *	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TOTAL PERCENT	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

* New item types on the 1994 examinations not found on the 1993 examinations.

BEST COPY AVAILABLE

TABLE 3B: THE VARIETY OF ITEM TYPES ON PUBLIC UNIVERSITY EXAMINATIONS, 1994

SKILL:	PUBLIC UNIVERSITIES									
ITEM TYPE	HITOTSU	HOKKAIDO	KYOTO	KYUSHU	NAGOYA	OSAKA	TOKYO	TORITSU	TYOUFS	YOKOHAMA
FREQUENCIES										
READING/WRITING:										
MULTIPLE-CHOICE	3	11	0	3	12	5	7	0	12	21
TRUE-FALSE *	0	0	0	0	0	0	0	0	0	0
REPHRASE/REORDER	0	0	0	0	0	0	0	0	1	0
FILL-IN	0	9	0	1	3	0	3	0	0	2
SHORT-ANSWER/ESSAY	5	1	0	6	5	2	3	6	10	2
TRANSLATION:										
TRANSLATE (E->J)	5	6	10	7	6	5	4	6	0	2
TRANSLATE (J->E)	2	4	2	3	4	2	1	5	0	4
LISTENING:										
TRUE-FALSE	0	0	0	0	0	0	0	0	0	0
MULTIPLE-CHOICE	0	0	0	0	0	0	0	0	5	0
FILL-IN	0	0	0	0	0	0	0	0	12	0
DICTATION	0	0	0	0	0	0	0	0	0	0
SHORT-ANSWER *	0	0	0	0	0	0	10	0	0	0
TOTAL NO. OF ITEMS	15	31	12	20	30	14	28	17	40	31
TIME ALLOWED	120	90	120	120	90	105	120	120	150	90
PERCENTAGES										
READING/WRITING:										
MULTIPLE-CHOICE	20.00	35.48	0.00	15.00	40.00	35.71	25.00	0.00	30.00	67.74
TRUE-FALSE *	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
REPHRASE/REORDER	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.50	0.00
FILL-IN	0.00	29.03	0.00	5.00	10.00	0.00	10.71	0.00	0.00	6.45
SHORT-ANSWER/ESSAY	33.33	3.23	0.00	30.00	16.67	14.29	10.71	35.29	25.00	6.45
TRANSLATION:										
TRANSLATE (E->J)	33.33	19.35	83.33	35.00	20.00	35.71	14.29	35.29	0.00	6.45
TRANSLATE (J->E)	13.33	12.90	16.67	15.00	13.33	14.29	3.57	29.41	0.00	12.90
LISTENING:										
TRUE-FALSE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MULTIPLE-CHOICE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	12.50	0.00
FILL-IN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	30.00	0.00
DICTATION	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SHORT-ANSWER *	0.00	0.00	0.00	0.00	0.00	0.00	35.71	0.00	0.00	0.00
TOTAL PERCENT	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

* New item types on the 1994 examinations not found on the 1993 examinations.

based on passages that the students had to read while others were relatively independent. All in all, with the various combinations of the factors just discussed, we found a tremendous variety of different types of items. However, even this is an oversimplification. In short, we were amazed at the variety of different types of multiple-choice items that the human mind can create.

We found similar variety within each of the other types of items shown in Tables 3A and 3B. We are not the first to notice this phenomenon. As Duke (1986) put it, ". . . written English tests in public schools can be quite intricate. They often require detailed knowledge of the techniques of the language, both written and oral" (p. 154).

One ramification of such intense variation in item types is that students are forced to change item types often within any given test. This means that new sets of directions (always in Japanese) are common. As a result, these examinations must assess the students' testwiseness, at least to some degree, in that they measure the students' abilities to handle varied and novel item types, read directions (in Japanese), and switch gears often. For years, tests in western countries have avoided such issues by keeping the item types similar within fairly large subtests. This practice is based on the belief that tests should assess the students' abilities in the content area (in this case, English) rather than their abilities to take tests.

Test length. Another fact that emerges from Tables 3A and 3B is that these examinations varied considerably in terms of sheer length, that is, in the time allowed and the numbers of items involved in each examination. The amount of time allowed ranged from 80 minutes for KYOTOUFS to 150 minutes for TYOUFS. The numbers of items involved varied from lows of eight, twelve, and fourteen at KEIO, KYOTO, and OSAKA, respectively, to highs of 79, and 80 at KANGAI and SOPHIA, respectively. Of course, such differences in numbers of items are related at least in part to the item types involved. For instance, the KEIO examination was made up of only 8 items, but half of those items were translation items (three

English to Japanese and one Japanese to English), which are probably more involved tasks than multiple-choice items like those which dominate the SOPHIA examination.

Nonetheless, test length, in terms of numbers of items, is an important consideration, which can directly affect the reliability of an examination. As a result, examinations of this importance in the West, like the TOEFL examination, are typically much longer in terms of both numbers of items and time allotted.

4. *Are there differences in the types of items and test lengths found in private and public university examinations in 1993 and 1994?*

Table 4 summarizes the 1993 information presented in Brown and Yamashita (1995) and the 1994 information presented in the previous section for the private and public universities grouped together, along with the *center* examination statistics and the totals for all 21 universities in each year taken together. Remember, the figures in this table are mostly averages.

On the whole, most of the same types of items were used at least sometimes in 1993 and 1994 in both the private and public universities. However, there were several notable exceptions to that general statement. Listening true-false items were only used in the private universities in 1993 and listening dictation items were only used in the private university exams in 1993 and 1994, while listening fill-in items were only used in the public universities in 1993 and 1994. In addition, two new types of items were used in 1994 that had not appeared on the 1993 examinations: true-false reading/writing items, which were used only in the 1994 private university exams, and short-answer listening items were where used only in the 1994 public university exams. Clearly, in both 1993 and 1994, the *center* examination used predominantly multiple-choice items with a few rephrase/reorder types of items providing the only variety.

Notice how the balance maintained among the various item types is very different on average when comparing the private and public universities. For example, on average,

TABLE 4: ITEM TYPE VARIETY SUMMARIZED BY UNIVERSITY TYPE *

SKILL: ITEM TYPE	1993 EXAMS PRIVATE	PUBLIC	CENTER	TOTAL	1994 EXAMS PRIVATE	PUBLIC	CENTER	TOTAL
FREQUENCIES								
READING/WRITING:								
MULTIPLE-CHOICE	32.10	3.60	48	19.29	33.70	7.40	55	22.19
TRUE-FALSE *	0.00	0.00	0	0.00	0.50	0.00	0	0.24
REPHRASE/REORDER	1.40	1.50	5	1.62	0.30	0.10	4	0.38
FILL-IN	5.70	3.10	0	4.19	6.50	1.80	0	3.95
SHORT-ANSWER/ESSAY	0.30	5.30	0	2.67	1.10	4.00	0	2.43
TRANSLATION:								
TRANSLATION (E->J)	1.40	4.20	0	2.67	1.20	5.10	0	3.00
TRANSLATION (J->E)	0.70	1.70	0	1.14	0.90	2.70	0	1.71
LISTENING:								
TRUE-FALSE	1.00	0.00	0	0.48	0.00	0.00	0	0.00
MULTIPLE-CHOICE	1.00	1.60	0	1.24	1.00	0.50	0	0.71
FILL-IN	0.00	4.00	0	1.90	0.00	1.20	0	0.57
DICTATION	0.10	0.00	0	0.05	0.10	0.00	0	0.05
SHORT-ANSWER *	0.00	0.00	0	0.00	0.00	1.00	0	0.48
TOTAL NO. OF ITEMS	43.70	25.00	53	35.24	45.30	23.80	59	35.71
TIME ALLOWED	93.50	112.50	80	103.00	95.50	112.50	80	102.86
PERCENTAGES								
READING/WRITING:								
MULTIPLE-CHOICE	63.55	13.98	90.57	41.23	62.31	26.89	93.22	46.92
TRUE-FALSE *	0.00	0.00	0.00	0.00	1.02	0.00	0.00	0.49
REPHRASE/REORDER	2.54	4.94	9.43	4.01	0.67	0.25	6.78	0.76
FILL-IN	16.82	12.97	0.00	14.19	15.62	6.12	0.00	10.35
SHORT-ANSWER/ESSAY	1.77	24.17	0.00	12.36	6.80	17.50	0.00	11.57
TRANSLATION:								
TRANSLATION (E->J)	6.64	18.41	0.00	11.93	6.06	28.28	0.00	16.35
TRANSLATION (J->E)	2.61	7.88	0.00	4.99	3.31	13.14	0.00	7.83
LISTENING:								
TRUE-FALSE	2.86	0.00	0.00	1.36	0.00	0.00	0.00	0.00
MULTIPLE-CHOICE	2.86	5.31	0.00	3.89	4.00	1.25	0.00	2.50
FILL-IN	0.00	12.35	0.00	5.88	0.00	3.00	0.00	1.43
DICTATION	0.36	0.00	0.00	0.17	0.20	0.00	0.00	0.10
SHORT-ANSWER *	0.00	0.00	0.00	0.00	0.00	3.57	0.00	1.70
TOTAL PERCENT	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

*All statistics for PRIVATE and PUBLIC universities as well as TOTAL are averages.

BEST COPY AVAILABLE

the private universities relied much more heavily on multiple-choice items in 1993 and 1994, while the public universities put a heavier emphasis on the short-answer/essay and translation types of items in both years. In addition, the private examinations had considerably more items on average in 1993 and 1994 than the public examinations. The time allotted was also different, with the average private university examination being considerably shorter in both years than the average public university.

Thus a student who wants a relatively quick examination with a fairly large proportion of multiple-choice items should focus on taking private university examinations, or, even more so, on the *center* examination.

5. *What skills were measured on the 1993 and 1994 English language entrance examinations?*

As we explained in the *Definitions* section above, different skills are necessary to answer discrete-point test items and integrative ones, as is the case with receptive test items and productive ones. We also argued that translation is probably too difficult, demanding, and specialized a skill to require of students who have only studied a language through junior and senior high school. Tables 5A and 5B presents the results for these different item categories. These two tables represent the private and public universities, respectively, and the tables are organized just like those which preceded.

Consider first the comparisons shown in Tables 5A and 5B between discrete-point, integrative, and translation items. Among the private universities (looking at the percents shown in Table 5A), the discrete-point items dominate all of the tests except for KEIO (which is 50 percent translation and 50 percent integrative). However, in Table 5B, the public universities generally appear to use fewer discrete-point items and the use of discrete-point, integrative, and translation appears to vary more.

In the same tables, receptive, productive, and translation items are also compared. In Table 5A the private universities have about the same pattern that occurred in the discrete-point results, that is, the receptive items dominated, probably because discrete-point items tend to be receptive (though some items can be discrete-point and productive). In Table 5B, the proportions of receptive items are generally less important in the public university tests and the proportions of receptive, productive, and

TABLE 5A: CATEGORIES OF ITEM TYPES ON PRIVATE UNIVERSITY EXAMINATIONS, 1994

ITEM CATEGORY	PRIVATE UNIVERSITIES AOYAMA	DOSHISHA	KEIO	KANGAI	KANSAI	KYOTOUFS	RIKKYO	SOPHIA	TSUDA	WASEDA
FREQUENCIES										
DISCRETE-POINT	20	43	0	79	43	43	37	80	44	31
INTEGRATIVE	1	0	4	0	3	1	1	0	1	1
TRANSLATION	4	2	4	0	2	4	1	0	4	0
NO. OF ITEMS	25	45	8	79	48	48	39	80	49	32
RECEPTIVE	20	37	0	79	43	43	23	80	13	17
PRODUCTIVE	1	6	4	0	3	1	15	0	32	15
TRANSLATION	4	2	4	0	2	4	1	0	4	0
NO. OF ITEMS	25	45	8	79	48	48	39	80	49	32
PASSAGE DEPENDENT	22	30	7	14	33	20	21	80	15	16
PASSAGE INDEPENDENT	3	15	1	65	15	28	18	0	34	16
NO. OF ITEMS	25	45	8	79	48	48	39	80	49	32
PERCENTAGES										
DISCRETE-POINT	80.00	95.56	0.00	100.00	89.58	89.59	94.88	100.00	89.80	96.87
INTEGGRATIVE	4.00	0.00	50.00	0.00	6.25	2.08	2.56	0.00	2.04	3.13
TRANSLATION	16.00	4.44	50.00	0.00	4.17	8.33	2.56	0.00	8.16	0.00
TOTAL % OF ITEMS	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
RECEPTIVE	80.00	82.23	0.00	100.00	89.58	89.59	58.98	100.00	26.53	53.12
PRODUCTIVE	4.00	13.33	50.00	0.00	6.25	2.08	38.46	0.00	65.31	46.88
TRANSLATION	16.00	4.44	50.00	0.00	4.17	8.33	2.56	0.00	8.16	0.00
TOTAL % OF ITEMS	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
PASSAGE DEPENDENT	88.00	66.67	87.50	17.72	68.75	41.67	53.85	100.00	30.61	50.00
PASSAGE INDEPENDENT	12.00	33.33	12.50	82.28	31.25	58.33	46.15	0.00	69.39	50.00
TOTAL % OF ITEMS	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

translation items vary more from university to university than among the private universities.

Most of the universities in Tables 5A and 5B rely heavily on items that are based on reading or listening passages, KANGAI and TSUDA being the only universities that based less than 40 percent of their items on passages of some sort.

Table 6 compares the 1993 and 1994 averages for private and public universities with the *center* examination and the total for all 21 universities. This table re-enforces the observations made above that the private university examinations are much more prone to using discrete-point receptive items than the public ones, and that observation is true for both 1993 and 1994. However, while the public university examinations are laudably using more integrative items, they are also relying heavily on translation items.

The patterns in Table 6 for the receptive, productive, and translation items are similar. In neither 1993 or 1994 are students being required to use any extensive amounts of productive language—not much written language, and absolutely no spoken language.

In both years, the public universities appear to use a higher percent of passage-based items than the private universities, and both types of universities appear to be using more passage-based items in 1994 than in 1993. Note also that the *center* examination used a much smaller proportion of passage-based items in both years than did either the public or private universities.

The listening skill, heavily promoted in the Monbusho guidelines that were implemented in spring of 1993 (Monbusho, 1989), appears in only a few of the examinations. According to Brown and Yamashita (1995), in 1993, six out of the 21 universities included listening comprehension items: AOYAMA, TSUDA, HITOTSU, KYOTO, TOKYO, and TYOUFS. Careful inspection of Tables 3A and 3B reveals that, in 1994, only four of the examinations (out of 21) included listening comprehension items of some sort: AOYAMA, TSUDA, TOKYO, and TYOUFS.

Discussion

One question that remains is what the results of this study mean to teachers and students of EFL in Japan?

Readability. Readability analysis of the passages used on the 1994 entrance examinations (see Tables 1A, 1B, and 2) indicated that all of the private and public universities used at least

TABLE 5B: CATEGORIES OF ITEM TYPES ON PUBLIC UNIVERSITY EXAMINATIONS, 1994

	PUBLIC UNIVERSITIES									
ITEM CATEGORY	HITOTSU	HOKKAIDO	KYOTO	KYUSHU	NAGOYA	OSAKA	TOKYO	TORITSU	TYOUFS	YOKOHAMA
FREQUENCIES										
DISCRETE-POINT	3	20	0	4	15	5	10	0	30	23
INTEGRATIVE	5	1	0	6	5	2	13	6	10	2
TRANSLATION	7	10	12	10	10	7	5	11	0	6
TOTAL NO. OF ITEMS	15	31	12	20	30	14	28	17	40	31
RECEPTIVE	3	11	0	3	12	5	7	0	18	21
PRODUCTIVE	5	10	0	7	8	2	16	6	22	4
TRANSLATION	7	10	12	10	10	7	5	11	0	6
TOTAL NO. OF ITEMS	15	31	12	20	30	14	28	17	40	31
PASSAGE DEPENDENT	13	14	10	14	16	13	21	12	26	13
PASSAGE INDEPEND.	2	17	2	6	14	1	7	5	14	18
TOTAL NO. OF ITEMS	15	31	12	20	30	14	28	17	40	31
PERCENTAGES										
DISCRETE-POINT	20.00	64.52	0.00	20.00	50.00	35.71	35.71	0.00	75.00	74.19
INTEGRATIVE	33.33	3.23	0.00	30.00	16.67	14.29	46.43	35.29	25.00	6.46
TRANSLATION	46.67	32.25	100.00	50.00	33.33	50.00	17.86	64.71	0.00	19.35
TOTAL % OF ITEMS	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
RECEPTIVE	20.00	35.48	0.00	15.00	40.00	35.71	25.00	0.00	45.00	67.75
PRODUCTIVE	33.33	32.26	0.00	35.00	26.67	14.29	57.14	35.29	55.00	12.90
TRANSLATION	46.67	32.26	100.00	50.00	33.33	50.00	17.86	64.71	0.00	19.35
TOTAL % OF ITEMS	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
PASSAGE DEPENDENT	86.67	45.16	83.33	70.00	53.33	92.86	75.00	70.59	65.00	41.94
PASSAGE INDEPEND.	13.33	54.84	16.67	30.00	46.67	7.14	25.00	29.41	35.00	58.06
TOTAL % OF ITEMS	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

TABLE 6: CATEGORIES OF ITEM TYPES SUMMARIZED BY UNIVERSITY TYPE *

ITEM CATEGORY	1993 EXAMS PRIVATE	1993 EXAMS PUBLIC	1993 EXAMS CENTER	1993 EXAMS TOTAL	1994 EXAMS PRIVATE	1994 EXAMS PUBLIC	1994 EXAMS CENTER	1994 EXAMS TOTAL
FREQUENCIES								
DISCRETE-POINT	39.80	12.30	48	27.10	42.00	11.00	59	28.05
INTEGRATIVE	1.80	6.80	5	4.33	1.20	5.00	0	2.95
TRANSLATION	2.10	5.90	0	3.81	2.10	7.80	0	4.71
TOTAL NO. OF ITEMS	43.70	25.00	53	35.24	45.30	23.80	59	35.71
RECEPTIVE	41.20	9.80	53	26.81	35.50	8.00	59	23.52
PRODUCTIVE	0.40	9.30	0	4.62	7.70	8.00	0	7.48
TRANSLATION	2.10	5.90	0	3.81	2.10	7.80	0	4.71
TOTAL NO. OF ITEMS	43.70	25.00	53	35.24	45.30	23.80	59	35.71
PASSAGE DEPENDENT	15.30	12.30	15	13.86	25.80	15.20	14	20.19
PASSAGE INDEPENDENT	28.40	12.70	38	21.38	19.50	8.60	45	15.52
TOTAL NO. OF ITEMS	43.70	25.00	53	35.24	45.30	23.80	59	35.71
PERCENTAGES								
DISCRETE-POINT	86.08	44.61	90.57	66.54	83.63	37.51	100.00	62.45
INTEGRATIVE	4.68	29.10	9.43	16.54	7.00	21.07	0.00	13.37
TRANSLATION	9.24	26.29	0.00	16.92	9.37	41.42	0.00	24.18
TOTAL % OF ITEMS	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
RECEPTIVE	88.63	37.19	100.00	64.68	68.00	28.39	100.00	50.66
PRODUCTIVE	2.13	36.52	0.00	18.40	22.63	30.19	0.00	25.16
TRANSLATION	9.24	26.29	0.00	16.92	9.37	41.42	0.00	24.18
TOTAL % OF ITEMS	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
PASSAGE DEPENDENT	41.71	52.66	28.30	46.29	60.48	68.39	23.73	62.49
PASSAGE INDEPENDENT	58.29	47.34	71.70	53.71	39.52	31.61	76.27	37.51
TOTAL % OF ITEMS	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

*All statistics for PRIVATE and PUBLIC universities as well as TOTAL are averages.

one passage and that some used as many as seven passages. The passages ranged considerably in length and readability. As with the 1993 examinations, the Flesch-Kincaid and Fog readability indexes indicated that the passages would be appropriate for native speakers ranging from eighth grade (about 13 years old) to thirteenth or even late fifteenth grade level (university age) at one university. Especially the passages at the upper end of these scales should be considered very difficult reading material for EFL students just finishing high school. In 1993, four universities had university level reading passages according to the Fog index (KEIO, KYOTO, TOKYO, and YOKOHAMA), while, in 1994, six universities had passages at this level—three private (DOSHISHA, KEIO, and TSUDA) and three public (NAGOYA, OSAKA, and YOKOHAMA). Table 2 indicates that, in 1993, the public universities had more difficult passages than the private universities, but, in 1994, the reverse was true. However, on the whole, the average level of difficulty across all of the universities did not change much between 1993 and 1994 as indicated by the TOTAL averages.

For those teaching EFL in Japan, the 1994 readability results (like the 1993 results) indicate that, by the end of their studies, university-bound high school students would benefit from learning to read relatively difficult university level passages with good comprehension.

Item types. The 1993 analysis of the item types (see Tables 3A, 3B, and 4) revealed that the reading/writing items included multiple-choice, rephrasing or reordering, fill-in the blank, and sort answer or essays, while the translations were both English to Japanese and Japanese to English and the listening items were true-false, multiple-choice, fill-in, and dictations. The same item types were used in 1994 with the addition of true-false items in the reading/writing category and short-answer items in the listening category. Thus university-bound EFL students should be equipped to deal successfully with these types of items, and with the considerable variation in formats that they represent. In short, we should probably prepare students for considerable variety in item types and frequent changes in the instructions/directions that go with those item types.

In 1993, DOSHISHA, KANGAI, KANSAI, KYOTOUFS, SOPHIA, TSUDA, and WASEDA used 50 percent of more multiple-choice items, while, in 1994, DOSHISHA, KANGAI, KANSAI, KYOTOUFS, RIKKYO, SOPHIA, WASEDA, and YOKOHAMA did so. In 1993, only KEIO and TORITSU used more than 50 percent translation items, while, in 1994, KEIO, KYOTO, KYUSHU, OSAKA, and TORITSU did so. Students should probably be advised of the types of items that have predominated in the last two years at whatever universities they want to enter.

It is worth noting the considerable differences in the test lengths both in time and in numbers of items. This fact can be important to individual test takers, but can also have important policy level implications because of the direct relationship between test length and test reliability (i.e., longer tests tend to be more reliable than shorter tests if all other factors are held constant; see Brown, 1995).

The skills analysis (Tables 5A, 5B, and 6) indicated that generally the proportions of discrete-point and receptive items were very high for the private university examinations, but less important in the public university tests, and that the proportions of receptive, productive, and translation items varied more from university to university among the public universities than among the private ones. It was also noted (from Tables 3A and 3B) that six of the 1993 examinations (AOYAMA, TSUDA, HITOTSU, KYOTO, TOKYO, and TYOUFS) out of these 21 included listening comprehension items, while in 1994 the number had dropped to four (AOYAMA, TSUDA, TOKYO, and TYOUFS).

Conclusions

Generally, the items on the tests that we analyzed in both 1993 and 1994 were reasonably well written with very few malapropisms, typographical errors, unintentional grammatical errors, etc. However, that does not mean that the tests are perfect. In fact, it is safe to say that no language test is ever perfect. The remainder of this chapter will discuss some of the problems that we see with the examinations that we analyzed.

Many of the items on these examinations were based on reading passages, and as we pointed out earlier, a number of the passages were difficult. Hence, the ability of a given student to answer these questions will depend to some degree on their ability to deal with relatively high level language that is perhaps above the level of the simplified texts that are often used for pedagogical purposes in Japan.

In addition, the ability to answer many of the items may depend on the students' knowledge of the particular topics involved in the passages. In other words, chance knowledge of a particular topic and its vocabulary could be helping some students to be accepted into a particular university, while other students, who were not lucky enough to have that chance knowledge, are excluded.

Before analyzing our 1993 results, we were skeptical of the value of the Japanese obsession with sending their children to a *juku* (i.e., a cram school) or *yobikô* (i.e., a test-coaching school) to prepare for major entrance examinations. However, based on both the 1993 and 1994 results, we now believe that such preparation may be advisable. For one thing, there is considerable variation in the types of items used on these tests, especially in the public university examinations. Such variation means that students are often reading new directions/instructions and shifting gears in terms of the kinds of items that they are answering. The ability to understand directions and shift gears on an examination is part of what is known as testwiseness, and testwiseness is one of the issues dealt with in a typical *juku* or *yobikô*. Whether we like it or not, given the very competitive nature of these examinations, testwiseness, or the ability to take tests in general, may be as important or even more important than the students' actual proficiency in English.

Teachers should also recognize the relationship between the item types used on university entrance examinations and the pedagogical choices that they make in their classrooms. In 1993 and 1994, the private universities predominantly used discrete-point receptive items. This means that in effect they were endorsing a discrete-point receptive view of language teaching. Many of the public universities were predominantly using translation items, which means that they were tacitly endorsing the use of translation as a communication strategy.

Students who are preparing for examinations of one type or the other, may quite reasonably want to focus on discrete grammar points, or translation tasks, and have very little interest in the communicative language learning or task-based learning that a particular teacher may be using if they do not see any direct relationship between what the teacher is doing and the examination that they must eventually take. This effect is called the washback effect, i.e., the effect of a test on the pedagogy associated with it (for more on washback effect, see Gates in Chapter 11 of this book). The present nature of the university entrance examinations appears to lead to a negative washback effect on efforts to employ modern language teaching methods.

The predominant North American examination, the TOEFL, is also discrete-point in nature, and it has been argued for years that it has a negative washback effect on modern language teaching methods. In response, Educational Testing Service has initiated the TOEFL 2000 project, which is aimed at changing the TOEFL into a more communicative and task-oriented examination. Perhaps, Japanese universities should begin to change their examinations in similar ways so that their washback effect can become a positive and progressive force for change in language teaching in Japan.

A contradiction has also developed between what is included on these university entrance examinations and the Monbusho (1989) guidelines implemented in April 1993 for junior and senior high school English teaching. The guidelines advocate the addition of listening and/or speaking to the curriculum, but our analysis indicates that only six universities in 1993 and four in 1994 included even a listening component. For students, taking the examinations without a listening component, there is a distinct negative washback effect on their desire to improve their listening or speaking abilities. Perhaps the structure and nature of the entrance examination items should change over the next several years to reflect the Monbusho's new emphasis on oral/aural skills.

The universities in this study openly allow for publication of their examinations, which is beneficial because it allows for public scrutiny of the items that were used. However, these universities should also be held responsible for defending the reliability, validity, and practicality of their tests. We have reason to believe that the entrance examinations in both the 1993 and 1994 studies are weak in all three areas. For instance, many of the tests are relatively short, which may pose a threat to their reliability. In addition, the types of items involved are not consonant with current language teaching theory and practice, which is a serious threat to validity. Finally, the tests do not appear to be 100 percent practical because the item types change frequently and many items are passage and topic dependent. Moreover, none of the universities that we are aware of do any of the statistical analyses of reliability and validity that are standard practice on major examinations (even at individual universities) in the United States.

How can these problems be avoided? As we did in the previous study of the 1993 examinations, we strongly recommend that Japanese universities follow the *Standards for Educational and Psychological Testing* (CDSEPT, 1985) or adapt those standards to create a set that will be acceptable in Japan. In the United States, tests only became consistently reliable and valid when students filed law suits against the various institutions that developed tests. The *Standards for Educational and Psychological Testing* then appeared and began to be applied to tests.

In addition, *Mental Measurements Yearbook* (e.g., Kramer & Conoley, 1992) provides periodic critical reviews of all tests published in the United States. Both the *Standards* and *MMY* help to keep test developers honest. Similar institutions in Japan might help improve the reliability and validity of the entrance examinations used in Japanese universities.

Perhaps the single most important fact about these entrance examinations is that the results are used to make very important decisions about students' lives. As such, they must be of the highest quality if they are to be fair to the students. In fact, the entrance examinations in Japan are far too important to be left entirely up to groups of individual test designers. Even university professors must be made accountable for the important admissions decisions that they are making, because those decisions so profoundly affect young Japanese lives.

Notes:

[1] We would like to thank Ryutaro Yamashita for his help in entering the reading passages into computer files.

References

Amano, I. (1990). *Education and examination in modern Japan*. Tokyo: Tokyo University Press. (Translated into English by K. & F. Cummings)

Borland. (1991). *QuattroPro* (version 3.0). Scotts Valley, CA: Borland International.

Brown, J. D. (1993). Language testing hysteria in Japan. *The Language Teacher, 17*(12), 41-43.

Brown, J. D. (1995). *Testing in language programs*. New York: Prentice-Hall.

Brown, J. D., & Yamashita, S. O. (1995). English language entrance examinations at Japanese universities: What do we know about them? *JALT Journal, 17*(1), 7-30.

CDSEPT (Committee to Develop Standards for Educational and Psychological Testing). (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Cummings, W. K. (1980). *Education and equality in Japan*. Princeton, NJ: Princeton University.

Dore, R. P. (1990). Foreword to I. Amano, *Education and examination in modern Japan* (pp. vii-x). Tokyo: Tokyo University Press.

Duke, B. (1986). *The Japanese school: Lessons for industrial America*. NY: Praeger.

Flesch, R. (1974). *The art of readable writing*. New York: Harper & Row.

Frost, P. (1991). Examination Hell. In E. R. Beauchamp (Ed.), *Windows on Japanese education* (pp. 291-305). New York: Greenwood Press.

Fujita, H. (1991). Education policy dilemmas as historic constructions. In B. Finkelstein, A. E. Imamura, & J. J. Tobin (Eds.), *Transcending stereotypes: Discovering Japanese culture and education* (pp. 147-161). Yarmouth, Maine: Intercultural Press.

Horio, T. (1988). *Educational thought and ideology in modern Japan: State authority and intellectual freedom*. Tokyo: Tokyo University Press.

Kiefer, C. W. (1970). The psychological interdependence of family, school, and bureaucracy in Japan. *American Anthropologist 72*, 66-75.

Kitamura, K. (1991). Japan's dual educational structure. In B. Finkelstein, A. E. Imamura, & J. J. Tobin (Eds.), *Transcending stereotypes: Discovering Japanese culture and education* (pp. 162-166). Yarmouth, Maine: Intercultural Press.

Kôkô-Eigo Kenkyû. (1993a). *'93 kokukôritsu daigaku-hen: eigomondai no tetteiteki kenkyû*. Tokyo: Kenkyû-Sha. *public*

Kôkô-Eigo Kenkyû. (1993b). *'93 shiritsu daigaku-hen: eigomondai no tetteiteki kenkyû*. Tokyo: Kenkyû-Sha. *private*

Kramer, J. J., & Conoley, J. C. (1992). *The eleventh mental measurements yearbook*. Lincoln, NE: University of Nebraska.

Monbusho. (1989). *Chû gakkô gakushû shidô yôryô* (Course of Study for junior high school). Tokyo: Monbushô.

National Council on Education Reform (NCER). (1985). Report. In Gyôsei (Ed.), *Rinkyôshin to kyôikukaikakujiyôka kara koseishugi e* (From liberalization to putting an emphasis on individuality). Tokyo: Gyôsei.

Organization of Economic Cooperation and Development (OECD). (1971). *A review of education: Japan*. Paris: OECD.

Que Software. (1990). *RightWriter: Intelligent grammar checker* (version 4.0). Sarasota, FL: Que Software.

Rohlen, T. P. (1983). *Japan's high schools*. Berkeley, CA: University of California Press.

Singleton, J. (1991) The spirit of gambaru. In B. Finkelstein, A. E. Imamura, & J. J. Tobin (Eds.), *Transcending stereotypes: Discovering Japanese culture and education* (pp. 119-125). Yarmouth, Maine: Intercultural Press.

Shimahara, N. (1991) Examination rituals and group life. In B. Finkelstein, A. E. Imamura, & J. J. Tobin (Eds.), *Transcending stereotypes: Discovering Japanese culture and education* (pp. 126-134). Yarmouth, Maine: Intercultural

Press.

Tsukada, M. (1991). Student perspectives on *juku, yobiko*, and the examination system. In B. Finkelstein, A. E. Imamura, & J. J. Tobin (Eds.), *Transcending stereotypes: Discovering Japanese culture and education* (pp. 178-182). Yarmouth, Maine: Intercultural Press.

Tsuneyoshi, R. K. (1991). Reconciling equality and merit. (1991). In B. Finkelstein, A. E. Imamura, & J. J. Tobin (Eds.), *Transcending Stereotypes: Discovering Japanese culture and education* (pp. 167-177). Yarmouth, Maine: Intercultural Press.

Vogel, E. *Japan's new middle class*. Berkeley, CA: University of California Press.

White, M. (1987). *The Japanese educational challenge: A commitment to children*. NY: The Free Press.

WordPerfect. (1988). *WordPerfect* (version 5.1). Orem, UT: WordPerfect Corp.

Chapter 11

Exploiting Washback from Standardized Tests

SHAUN GATES
SHIGA WOMENS' JUNIOR COLLEGE

The influence of testing on teaching and learning is known as washback. We can look at washback from two angles. First, washback may be strong or weak. If washback is strong, students and teachers will tend to alter their classroom behaviors in order to achieve good marks in the test. In contrast, weak washback will have little or no effect in the classroom. Another way to look at washback is to ask whether it is positive or negative. At this stage, we could simply define washback as being positive if test and course objectives coincide. Negative washback occurs when these two sets of objectives differ. The different possible types of washback (from a test or part of a test) can be located on the following grid:

	Positive	Negative
Strong		
Weak		

Teachers might reasonably want to determine the type of washback that flows from a given test. I suspect the English tests I give my Japanese college students fall into the bottom left box. Washback is positive because both the course and test objectives stem from the communicative approach, but one reason it is weak is that I am limited in the rewards and sanctions I can attach to test outcomes.

In the situation outlined above, testing is done for the internal consumption of the school or college. This is not always the case. Other teachers may be involved in preparing students for a standardized test, which by its nature provides results with a much wider currency. This chapter has two purposes: it attempts to explain why washback from standardized tests is so strong, and it shows how teachers can exploit this washback. All of this is done with reference to the *Preliminary English Test* (PET, 1995) run by the University of Cambridge Local Exam Syndicate (UCLES).

Research into washback suggests that the phenomenon is more subtle than was at first thought (Alderson & Wall, 1993). For Weir, the strong washback that derives from communicative teaching and testing suggests that washback may be linked to construct validity (1990, p. 27). But while the exact nature and influence of washback still has to be established, its force is well-recognized. Consider what happens when teaching and testing objectives diverge markedly, a situation that can arise if teaching and testing fail to develop together.

Davies (1990, pp. 96-100) provides some

interesting examples of this problem. What he calls the problem of ***excessive conservatism*** occurs when progress in teaching is not matched by an equivalent advance in testing. The adoption of a communicative curriculum in a school may have little effect on students' communicative competence if their end-of-course test requires them to write a literature essay.

The opposite problem, ***unthinking radicalism***, is the result of trying to impose desirable change in the classroom through the examination system. Imagine a situation where high school students have traditionally been required to translate a previously unseen prose passage in order to get into university. Then a new exam is introduced which calls for students to take an oral interview and a listening test. Unless teachers are retrained, new materials written, and students given time to adjust, the whole exercise is likely to end in frustration and even failure.

You will probably recognize the force of washback from your own experience of tests, whether they are academic, sporting, or whatever. As I write this chapter, I am also preparing (in vain) for a Japanese language test. My preferred style of learning is to dip into a range of books and then try out my new knowledge on friends and students. Now, with past test papers as a guide, I spend my time trying to memorize a large number of verb endings and *kanji* compounds.

Factors Affecting Washback

The following list gives some of the factors which may influence the strength of washback: prestige, accuracy, transparency, utility, monopoly, anxiety, familiarity, and practicality. Each of these factors deserves to be considered in more detail.

Prestige. A test will have strong washback if, like the *Test of English as a Foreign Language* (TOEFL), it is associated with a reputable, well-known organization (Educational Testing Service in this case). However, it is worth remembering that prestige does not necessarily mean widespread recognition. A test in interpretation, for example, may have stronger washback for potential interpreters than a better-known but more general language proficiency test.

Accuracy. Since language tests are used by employers and colleges to select suitable candidates, the score users are likely to base their decisions on those tests which give high levels of reliability and accuracy. This will in turn be picked up by students and teachers, who will concentrate on the relevant tests.

Transparency. The closer a language test meets the final language needs of the student, the stronger the washback will be. Direct tests which closely resemble real-life language use should increase student motivation and hence the force of washback. Indirect tests are less transparent and may have weaker washback.

Utility. The more opportunities a passing score in a test offers, the stronger its washback. A high TOEFL score not only gives Japanese students the chance to study at American universities, it may also help them find a job in a Japanese company that has foreign dealings. And there are other reasons why Japanese students take this test in large numbers, as Brown has argued (1993). Similarly, a pass in the *Cambridge Certificate of Proficiency* (CPE) not only satisfies the English language requirements of British universities, but in some European countries, it also serves as a valid qualification to teach English in government schools.

Monopoly. The less competitors an exam has, the stronger its washback. One of the notable things about British EFL exams is the sheer number of them. In addition to the CPE, there are at least three other ways foreign students can satisfy a British university that their English abilities are adequate. They can show they have a pass in British high school exams, in the IELTS exam, or (at some universities) in the TOEFL. Since students can choose the test which suits their situation and inclinations, the washback of each test is diluted.

Anxiety. Naturally, a test which puts excessive stress on the student will have strong washback. Whatever the advantages of direct tests, we should accept that some tasks such as participating in a role play or writing an expository essay may deter students from taking these tests. Indirect tests can cause stress too (particularly, if students don't know the answers!), but in this case, the stress can to some extent be controlled by the student; they can take a short break, guess the answer, or move on to an easier section. The more important the exam is in the student's eyes, the greater the level of anxiety is likely to be. Anxiety is also linked to the next factor.

Practicality. Tests which are convenient to sit, are held frequently, and are economical and short (without compromising their validity and reliability) will have stronger washback than tests that don't have these advantages. Other practicality issues would include the availability of tutors, published study guides, and practice tests.

Assuming that the strength of these factors varies between tests, it should be clear that standardized tests like the well-known English proficiency exams exhibit strong washback. The elements of a test that would appeal most to test takers––prestige, utility, and accuracy—are the washback factors that predominate in standardized tests. Furthermore, the standardized English tests that originate in the United States and Britain have a virtual monopoly in determining whether a foreign student's English is adequate for study on degree and vocational courses.

Even practicality, one of the strengths of school tests has been undermined by the advance of international English exams. For Japanese students, at least, tests exams are affordable and are held frequently throughout Japan. Role play and interviews may be familiar as classroom activities, but many Japanese students will be unfamiliar with their use as test instruments. As a result, the use of role play or interview procedures in communicative tests will probably lead to weak washback, although this may be offset somewhat by the transparency of these techniques.

In contrast, since most English tests in Japanese schools rely heavily on multiple-choice questions, the washback for TOEFL takers should be strong. Common sense suggests that the anxiety produced from a standardized test and a class test will be negative, but it will probably be much stronger in the former.

Exploiting Washback

Given that standardized tests have a strong influence on learning, teachers might ask how they can exploit this force to improve student performance. To answer that question, I must return to the definition of positive washback. It was stated above that positive washback happens when course and test objectives overlap. But what should those objectives be? There are two good reasons why I feel they should be communicative ones. First, the nature of language proficiency seems best captured by models that in addition to a knowledge of grammar also incorporate the instrumental role of language and the effect of context. Bachman's framework of **communicative language ability** represents one of the most comprehensive explanations along these lines and is based on a decade of model building and empirical research (1990, pp. 81-110). From a different perspective, Weir put forward a three- part framework to assess each of the four language skills (1993, pp. 28-29). While his framework focuses on performance rather than competence, it also argues for seeing language use as goal oriented and constrained by a range of conditions.

Second, communicative objectives fit in with most students' reasons for learning languages. Ask your students why they are learning English, and you will receive a variety of answers including **for travel, to study abroad,** and **to talk to foreigners.** All of these reasons for studying English, however, spring from a common desire to use English rather than dissect it.

There is a third and admittedly weak reason for choosing communicative objectives—default. In their choice of class textbooks and activities, many teachers have already instilled into their students the importance of communicative objectives.

Teachers who are responsible for their own courses and tests can achieve positive washback fairly easily. If they choose course and test objectives on the basis that they share the same orientation, there should be no tension between exam work and learning needs. Students will realize that class work gives both practice for the test and preparation for using English in authentic situations. Hughes gives useful advice on how to achieve beneficial backwash in class achievement tests (1989, pp. 44-47).

Unfortunately for the teacher who has students preparing for a standardized test, matters are not so clear cut. Individual teachers cannot influence the selection of content or the test techniques used in these standardized tests. The solution would seem to involve giving the students practice in all sections of the test irrespective of whether they lead to positive washback or not. The problem with this response is that it takes no account of the limited time available for learning English at most schools and colleges. A more efficient way to exploit these tests is to select only those items for classroom practice which meet the students' communicative goals. The remaining items can then be set for homework. An example of how this works is given for the UCLES *Preliminary English Test* (PET, 1995).

Table 1 *PET Item Specifications*

Q. SKILLS Objective	Text Type	Test Format
Reading Skills		
1. Understanding lexis	Public notice or sign	Multiple-choice
2. Understanding grammar and discourse	Short article, letter	Gap fill
3. Reading for detail	Adverts, brochure	Matching
4. Scanning	Brochures, notices	Matching
5. Reading for gist	Review, advert	Matching, or short answer
Writing Skills		
6. Grammatical accuracy	Five sentences	Sentence rewriting
7. Expressing	Form, note, message	Directed writing
8. Narrating	Letter, note	Free writing
Listening Skills		
9. Listening for specific information	Short utterance(s)	Multiple-choice
10. Listening for information	Factual report	Multiple-choice
11. Listening for main ideas	Factual report	Multiple-choice
12. Listening for gist	Conversation	True-false, or Yes/No
Speaking Skills (*An oral interview is used to test speaking subskills*)		
I. Social conversation–personal information.		
II. Role play–involving requests, directions, etc.		
III. Picture description–identification, comparison.		
IV. Discussion–likes, dislikes, personal experiences.		

An Example of Exploitation

The PET is aimed at students whose English level is pre-intermediate. The test is divided into the four traditional skills, but emphasis throughout the test is on being able to use English in real life situations. Table 1 shows each question number, the skill tested, the type of text used, and the test format involved.

As can be seen in the table, the PET contains a variety of authentic texts and is closely tied to language needs, making it a suitable test for students following a communicative course at this level. Washback should thus be positive. It should also be strong since the test, the product of a well-known exam board, would be rated highly for prestige, accuracy, and utility. However, exploiting a test in the classroom is not always straightforward even if, like the PET, it has strong positive washback.

Students preparing for this test at the British Council in Kyoto can take a special course in which they attend classes over 12 weeks. Because class time is limited to two hours per week, there is pressure to cover a number of the shorter questions particularly from the speaking and listening sections rather than one or two questions from the reading and writing sections. (When I use PET practice tests with my university students, there is even less time for test preparation.)

To compound the problem, the speaking and listening sections lend themselves to class work, naturally fitting in with student demands for more conversation practice. Positive washback from these skills, which can be easily translated into classroom activities, may drown out the much weaker washback coming from reading and writing sections. Writing in particular is a worry. Although writing, like the other sections, is given equal weighting in the marking scheme, this is probably offset by the great difficulty composition seems to cause Japanese students. Anecdotal evidence from the International English Language Testing Service (IELTS) test suggests that the writing section is often seen as the biggest obstacle in a test.

One way to remedy the problem is to set the bulk of the reading questions for homework and get the students to do one major writing exercise in class Fortunately, question 7 in Table 1 (on directed writing) can often be integrated with student requests for oral practice. For example, instead of filling in a form by himself, a student can interview his partner and complete the form with the partner's details. Question 8, the free writing question, which often requires the student to write a letter to an imaginary friend, will be more relevant if the student writes to a classmate, who then provides spoken or written feedback.

When dealing with the interview, the aim must be to give students practice but not let this section siphon off a disproportionate amount of class time. Mock interviews are quite common in many classrooms, and students will probably be familiar with the exchange of social formulas (number I. in Table 1). However, the section which asks students to comment on a picture (III.) may be novel, and students will almost certainly need practice to build up their skills and confidence for the discussion (IV.).

Failure to understand the spoken language can be very demotivating for students, a feeling which will be reinforced in an exam where 25 percent of the marks are allocated to the listening section. Obviously, class time must be spent on listening practice, but at the same time, we need to remember that the listening skill demands intensive concentration from the student. In the exam, the listening section takes half an hour, but it would probably be counterproductive to cover all the listening questions in a two-hour lesson. The aim should be to cover one or two questions in detail, and where possible, integrate the listening with other activities in the lesson. If practice texts are available, the teacher and students can read through the tapescripts picking out the students' weak points for remedial study.

Conclusion

Though it may not be possible to quantify the force of individual washback factors in a standardized test, their combination will probably produce strong overall washback. The teacher's role should be to ensure that the washback is also positive. Certainly, students need to practice questions before a test, but in the classroom, this practice should be designed to meet their broader language needs. As the discussion of the PET indicated, tests which have the potential to create positive washback come under pressure from limited class time and the students' preferences for some question types over others. Given the bother this creates, the temptation is for the teacher not to tell the students too much about the test. This strategy is unlikely to be successful. Most test takers soon find out the details of their test, and they will resent doing class work which they see as irrelevant to improving test performance. It is better to accept the existence of washback and harness it, much the same way an expert in *aikido* exploits the force of an opponent. With luck, the outcome for test takers and their teachers will be less painful.

References

Alderson, J. C., & Wall, D. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*(1), 41-69.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Brown, J. D. (1993). Language testing hysteria in Japan? *The Language Teacher, 17*(12), 43-43.

Davies, A. (1990). *Principles of language testing.* Oxford: Blackwell.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University Press.

PET. (1995). *A guide to the Preliminary English Test.* Cambridge: UCLES.

Weir, C. J. (1990). *Communicative language testing.* Hemel Hempstead, UK: Prentice-Hall.

Weir, C. J. (1993). *Understanding and developing language tests.* Hemel Hempstead, UK: Prentice-Hall.

Section IV

Oral Proficiency Testing

Chapter 12

Testing Oral Ability: ILR and ACTFL Oral Proficiency Interviews

HIROTO NAGATA
MEIKAI UNIVERSITY
&
OVERSEAS TRAINING CENTER AT PANASONIC, TOKYO

Assessing students' oral proficiency is a perplexing problem for many language teachers. We all know that pencil-and-paper tests are not valid measures of oral production, i.e., they cannot adequately appraise learners' abilities in the functional use of a foreign language, and we also know that oral interviews are best used for that purpose. However, we do not usually have enough time to participate in a training program to become a full-fledged, qualified interviewer. Nonetheless, with the Ministry of Education placing new emphasis on communicative goals in language teaching, the need for classroom teachers to be equipped with some measurement tools to evaluate students' oral proficiency is becoming more and more important.

This chapter presents two simple and easy-to-handle models called the Interagency Language Roundtable (ILR) oral proficiency interview (Lowe, 1982) and ACTFL oral proficiency interview (ACTFL, 1986). Both the ILR oral proficiency interview and ACTFL oral proficiency interview are derived from a similar test developed in the 1950s by the Foreign Service Institute of the United States State Department to assess the foreign language ability of United States Government agency staff. This Government rating scale was later modified for use with students in secondary school and college foreign language programs by the ACTFL and the Educational Testing Service.

In the first half of this chapter, the rating scales of both the ILR and its close derivative, ACTFL oral proficiency interview, will be described. The structure of an interview, and some example elicitation questions and question types will also be explained. In addition, information regarding the disadvantages of certain types of questions will be provided. In the second half of the chapter, some basic problems regarding oral interviews will be discussed in the light of Dos and Don'ts that interviewers should bear in mind. Advice and various suggestions concerning planning and conducting oral interviews will then follow.

The Rating Scales

An oral interview typically takes approximately 20 minutes in a face-to-face conversation between one interviewee and one or two interviewers. In the case of ILR ratings, on which the ACTFL oral proficiency interview is based, the conversations are rated on a scale of 0 (for no practical speaking proficiency) to 5 (for proficiency equivalent to that of an educated or well-informed native speaker). In between these 6 levels (0 to 5), are five 'plus' ratings (0+ to 4+), which indicate ability at almost the next higher level, so this makes 11 levels all together.

Table 1 (see Appendix) shows ILR guidelines at each level (ETS, 1982, pp. 34-36). Note that each description is broad enough to include both weaker and stronger performances over a significantly wide range. Interviewers are required to familiarize themselves with these criteria and have at their fingertips the characteristic features that distinguish each level. Ratings are determined by comparing the totality of a student's speaking performance to the descriptions at each level, i.e., no single instance of strength or weakness should determine the final rating.

As you can tell immediately, levels 3, 4, and 5 in the ILR scale are not really necessary for most purposes in dealing with second language learners in Japan. Most Japanese learners, unless they are well-educated bilinguals, recent returnees, or students exceptionally endowed with language learning ability, fall somewhere between 0+ and 2+, and very rarely 3, which is considered to be a highly proficient second language learner. Narrower and more precise descriptions are definitely needed here in Japan to discriminate among students who would score between levels 0 and 2+ on the ILR oral proficiency interview. This is where the ACTFL scale comes into the picture. Table 2 (see Appendix) displays the relationships between the ILR and ACTFL scale (ACTFL, 1989, 2-15). It should be clear that the ACTFL scale is a more detailed expansion of the ILR's lower levels, from 0 to 2+.

When rating, therefore, interviewers could first go from broader descriptions, placing the performance of an interviewee within the appropriate range —0 to 0+, 1 to 1+, and 2 to 2+ of the ILR scale, which corresponds to Novice, Intermediate, and Advanced of the ACTFL scale. Once the range has been determined, depending on the teacher's needs, she could proceed to refine the rating to reflect the sub-level descriptions of Low, Mid, or High. If the teacher only needs to roughly assign her students, broader descriptions, such as Novice, Intermediate, and Advanced will suffice. However, if teachers want to give remedial guidance to each of their students, for instance, more finely-tuned descriptions will be called for.

Table 3 (see Appendix) shows the proficiency descriptions of the ACTFL scale (ETS, 1982, i-iii.). Note that the difference between Novice Low and Novice Mid is one of quantity, and that between Intermediate Low and Mid is affected by both the quantity and quality of interviewees' performances.

The Structure of an Interview

As shown in Figure 1 (see Appendix), an interview has four phases: warm-up, level check, probes, and wind-down. Let us look at each phase in terms of its purpose, and see how you should conduct yourself in each.

The Warm-up (3-5 Minutes)

The main purpose of the warm-up is to put the interviewees at ease and open the way for further conversational exchanges. Therefore, this phase, as an opening to an interview, consists of social amenities and such simple conversation as introducing yourselves. The length varies depending on the interviewee's level of proficiency. Those who have been out of practice, for instance, might need time to get back to the language gradually, while others having daily contact with the language, or those with a high proficiency level may not need to be bothered with simple social rituals. The second, more important purpose of the warm-up phase is to get a preliminary indication of the interviewee's level, which is to be checked closely in the next phase, the level check.

The Level Check (8-10 Minutes)

The purpose of the level check is to find the highest level at which the interviewee can sustain speaking performance. In this phase, the interviewer should check a number of topics to see if the interviewee can perform consistently at the level in question. Once in a while, the level indication the interviewer obtained in the Warm-up phase will be faulty and the Level check might begin too low or too high. If this happens, interviewers can simply raise or bring down the level of the questions and resume the level check without making a big fuss over it. In order to find the interviewee's highest sustained level, fluency, accuracy, width of vocabulary, and content must be tested. Is the interviewee fluent enough to be at that certain level? Is her grammar accurate? How about her syntax? Can she perform the functions prescribed in the rating descriptions using suitable content? If the interviewee successfully passes this Level check, her performance provides a "floor" for the rating. The next phase, the Probes, attempts to find the "ceiling" of the interviewee's abilities.

The Probes (7-10 Minutes)

The purpose of this phase is to make sure the level the interviewer has been checking is the interviewee's highest sustained level. In order to determine this, the interviewer should take the interviewee one level above the present level several times, preferably, three to four times. While the Level check gives evidence of what an interviewee can do, the probes show what an interviewee cannot do. Therefore, if the interviewer's level check has been appropriate and the highest sustained level has been correctly established, probes will end up causing linguistic breakdown: a sharp drop in fluency, or accuracy (e.g., a dramatic increase in grammatical or syntactic errors). The interviewee may also confess to the interviewer that the limit (the "ceiling") has been reached by saying something like, "I don't know how to say it," or "It's difficult to say." If, on the other hand, the interviewer has carried out the level check at too low a level, the interviewee will consistently react and perform well on the probes. If that happens, the interviewer must begin the level check and Probes over again at a higher level, and continue the process until the interviewee's proficiency ceiling is established. It is imperative that assignment of the rating be done by comparing the totality of the interviewee's performance to the level descriptions, then, finding the one which most closely matches that performance. Under no conditions should one interviewee's performance be compared to that of another. This is important because comparing students with each other is a built-in tendency for many language teachers.

Wind-down (2-5 Minutes)

The purpose of the Wind-down is to return the interviewees to the level at which they perform best and let them leave the testing site with a sense of accomplishment.

Although most interviews follow the same general structure depicted above, these four phases might become indistinguishable at the very lowest levels, and a warm-up or wind-down might not be necessary at the very highest levels.

Useful Elicitation Questions and Question Types

Interview tests typically consist of a series of questions. Thus, it is no exaggeration to say that the success or failure of an interview depends largely on topic areas and types of questions asked. Table 4 (see Appendix) provides an overview of general topic areas and useful question types, as well as some information on the disadvantages of certain question types (adapted from ETS, 1982, pp. 43-57, 75-86).

Some Dos and Dont's

Since many interviewers are also teachers, they might increase the interviewees' discomfort, and thus decrease their desire to talk by bringing in certain behaviors deemed desirable in classrooms but not in an interview situation. For example, an overly helpful teacher who corrects students and finishes their sentences for them will not be a suitable interviewer. Teachers who have a tendency to fill in students' pauses by providing needed words, or filling in students' hesitations, cannot become good interviewers, either. Cutting an interviewee's answer short by giving long comments, expressing opinions excessively or too frequently also reduces an interviewee's chance to talk, thus depriving the interviewer of ratable material.

The following are some Don'ts to keep in mind during an interview (based on ETS, 1982, pp. 37-39.):

During the Warm-up:

1. Don't immediately launch into the Question and Answer Routine without the introductory "Hello, how are you?" Instead, make the interviewee feel at ease.
2. Don't ask the interviewee "Are you nervous?" or say "Gee,you really look nervous." Instead, suggest some positive soothing action: "Would you like to smoke, move your chair, make yourself comfortable?"
3. Don't immediately pose a difficult question involving hard or obscure grammar, idioms or vocabulary. Instead, talk about the weather, summer vacation, etc.

During the level check & the probes:

4. Don't look uninterested in what the interviewee is saying by looking constantly toward the floor, window or the clock. Instead, act interested in her experiences. Keep an eye contact, nod, smile, and be alert.
5. Don't play the role of authority: "I don't think you understood how Koreans feel: the truth is . . . " Instead, judge the interviewee on the language in which she expresses her thoughts.
6. Don't insist on a topic which is not the interviewee's field. She then might be given a lower rating than she deserves. Instead, follow up every clue which might lead to an area of interest. Probe for this as much as for levels.
7. Don't correct the interviewee's grammar during the interview. Instead, find out what the interviewee can do. Ask for clarification, if unclear.

A Few Other Do's and Don'ts

In his book *Testing for Language Teachers*, Hughes (1989, p. 105) presents 11 pieces of advice on planning and conducting oral tests. Three of these pieces of advice have not been dealt with in this chapter so far (time, number of testers, and affective factors) so they will be touched upon here.

Time. Hughes says it is unlikely that reliable information can be obtained in an interview of less than 15 minutes, while 30 minutes can probably provide all the information necessary for most purposes. However, he also contends that as part of a placement test, a five to ten minute interview should be sufficient to prevent gross errors in assigning student to classes. (For instance, by memorizing the simplified checklist shown in Table 5 (see Appendix), teachers can interview their incoming students, give ratings immediately after each interview, and assign a student to three to six classes of different levels (ACTFL's Novice, Intermediate, and Advanced, or ILR's 0, 0+, 1, 1+, 2, and 2+ levels) quite easily within a very short time (i.e., spending no more than five to ten minutes on each one).

Number of testers. Hughes recommends that a second tester be present for an interview. This is because of the difficulty of conducting an interview and keeping track of the candidate's performance. ACTFL (1989, p. 7-2) suggests that the interview be recorded so the rater can concentrate fully on rating the sample, and, in some difficult cases, seeking the opinion of another tester is recommended.

Affective factors. Hughes also warns against interviewers constantly reminding interviewees that they are being assessed. In particular, an interviewer should not be seen making notes about an interviewee's performance, either during the interview, or any other time. In the ILR and ACTFL oral proficiency interviews, taking notes is also prohibited. As mentioned in the previous section, in the ACTFL interviews, tape-recording is recommended. Hughes also recommends that transitions between topics be made as naturally as possible.

Conclusion

I have taken you quickly through the ILR proficiency rating scale and its close derivative, the ACTFL oral proficiency interview scale, including discussion of the structure of an interview, example elicitation questions and question types, and some dos and don'ts for interviewers.

Besides being a testing instrument, an interview provides one of the best ways to get to know your students. In the first place, it is fairly easy and quick to elicit the kinds of answers you need through an interview. It also gives you a thumbnail sketch of where your students stand on their way to acquiring procedural knowledge of the language. Face-to-face communication provides a good way to diagnose your students' weak points, and thus helps you offer remedial treatments rather than commenting on a sheet of paper. I have conducted ILR oral proficiency interviews for the past 15 years and kept records of the interview results of some interviewees to whom I gave the simplified checklist (see above) together with my comments. I also have records of those to whom I gave oral comments but not the checklist. Quite interestingly, those who were given the Checklist have improved their performances much more than those who were not. Having something (here, a printed checklist) that they can turn to and review/preview as a sort of beacon to sail across the sea of language acquisition appears to facilitate their self-directed learning.

Another benefit in having an oral interview implemented in a language program is that it makes many of the classroom activities acquire some new real-world reality. Many students in Japan still do not feel an imminent need to be capable of functioning in a foreign language. Classroom activities have very little reality if the students do not feel a need to use the language outside the classroom. To change the whole atmosphere of foreign language study from merely being an academic pursuit or expensive pastime or avocation, implementing an oral interview goes a long way.

For example, I use one part of each of my classroom hours to prepare for an interview test—or at least, that is the pretext I use, because in theory, you cannot really *prepare* for a proficiency interview. In my mind, this is a **five Ws and one H** activity, a question and response expansion exercise with some real-world meaning (see Nagata, 1993). It is a pair-work exercise in which student A first asks student B a yes/no question. Instead of answering that question, student B should convert that yes/no question into some Wh-questions without changing the general motive of the question. Student A, then answers the Wh-question, and continues the interaction as if playing the parts of an interviewer and interviewee interchangeably until they use up all the related topics they can talk about. One example might look something like the following:

Student A: Do you like Chinese food?
Student B: How do you like Chinese food? (or What do you think of Chinese food? or What kind of Chinese food do you like best?)
Student A: I like it very much, especially spring rolls.
Student B: What is a spring roll? (interaction continues)

Once in a while, I incorporate achievement elements in the interview. (Here, again, I am well aware that ILR and ACTFL oral interview tests are proficiency tests, not achievement tests.) Announcing, for instance, that a couple of S-1 Familiar Situations (see "Useful Elicitation Questions and Question Types") will be tested in the upcoming interview, though, gives students a strong incentive to enthusiastically practice those situational exchanges with their partners. It is not that they are practicing because those particular situational exchanges will be tested, but the practice itself now has real meaning for communication. Rote memorization of textbook dialogues, a common practice before implementation of the interview, is long gone from my classroom. Everybody knows memorized lines will not help much in the interview (and in real-world situations) because while dialogue situations are constant, real-world situations are always changing.

Tightly matching performance descriptions with the course entrance and exit requirements of a language program also helps learners visualize a clearly defined goal at each level. ILR and ACTFL oral proficiency interview ratings can be used profitably in that direction, too.

Teachers who have no time to worry about participating in an interview training program at this time can still benefit from the simplified checklist provided above. By simply listening to your students' interaction, you will not only be able to assign them to different level classes, but also be able to give them remedial guidance as well. For example, if one young woman can communicate simply through questions and answers, but cannot sustain control of past and future tenses or narrate and describe well, she is a 1 (on the ILR scale) or Intermediate (on the ACTFL scale). What she needs is the ability (a) to sustain control in the past, present, and future tense, (b) to narrate and describe, (c) to provide real conversational exchange (not just through questions and answers), and (d) to function in survival situations with a complication.

If we assume that the objective of teaching spoken language is development of the ability to interact successfully in the language, and that this ability involves not only comprehension but production, oral proficiency interview skills are a must for language teachers. I hope this article will help teachers in Japan place more emphasis on the oral production ability of their students.

References

ACTFL (American Council on Teaching of Foreign Languages). (1986). *ACTFL proficiency guidelines.* New York: American Council on Teaching of Foreign Languages.

ACTFL (American Council on Teaching of Foreign Languages). (1989). *ACTFL testers training manual.* New York: American Council on Teaching of Foreign Languages.

ETS (Educational Testing Service). (1982). ETS oral proficiency testing manual. Princeton NJ: Educational Testing Service.

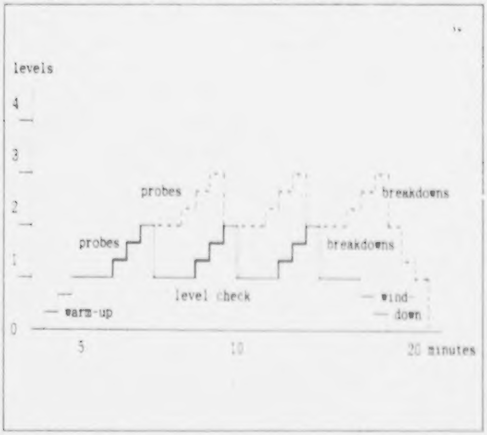
Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University.

Lowe, P. Jr. (1982). *ILR handbook on oral interview testing.* Washington, DC: DLI/LS Oral Interview Project.

Nagata, H. (1993). How to start a conversation: Small talk and praises. *The English Teachers' Magazine, 41*(12), 60-63.

Appendix: Figures and Tables

Figure 1. *The structure of an interview*



Level checks:	To find the candidates's highest sustained level.
Probes: (━━━)	To take the candidate one rank above (to make sure that the level the interviewer has been checking is really the interviewee's highest sustained level).
Breakdowns: (═══)	A sharp drop in fluency A sudden groping for words. A dramatic increase in grammatical errors.

Table 1. *ILR guidelines (ETS, 1982, pp. 34-36)*

Level 0: No Practical Proficiency

Speaking: The examinee has no practical speaking proficiency. May have a few isolated words and phrases which are of no practical use.

Understanding: The examinee understands some isolated words and phrases, but is unable to participate even in a very simple conversation.

Level 1: Survival Proficiency

Speaking—Subject matter: The examinee has the minimum proficiency for survival on a day-to-day basis in the target country, i.e., functions in simple question-and-answer situations. Knows enough at this level to satisfy ordinary courtesy requirements. Able to ask and answer questions relating to situations of simple daily life and routine travel abroad. The examinee is also able to handle requests for services such as renting a hotel room and ordering a simple meal.

Speaking—Quality: The examinee at this level normally makes errors even in structures which are quite simple and common. Vocabulary is limited to the type of situations mentioned above, and even in these situations he or she sometimes uses the wrong word. Although pronunciation may be poor, he or she makes the minimum contrastive distinctions, including stress, intonation and tone patterns necessary to be understood.

Understanding: The examinee is able to understand simple questions and statements relating to simple transactions involved in situations of daily life and independent travel abroad, allowing for slowed speech with considerable repetition or paraphrasing.

Level 2: Limited Working Proficiency

Speaking—Subject matter: The examinee is able to talk in some detail about concrete subjects such as own personal and educational background, family, travel experiences, recreational activities, and familiar places.

Speaking—Quality: The examinee has enough control of the morphology of the language (in inflected languages), and of the most frequently used syntactical structures. Although vocabulary is sufficient to talk with confidence about the type of topics described above, the limited vocabulary fairly often reduces the examinee to verbal groping, or to momentary silence. A foreign intonation and rhythm may still be dominant.

Understanding: The examinee is able to comprehend questions and statements relating to common social topics, when the language is spoken at normal conversational speed. Can get the gist of casual conversations with educated or well-informed native speakers talking about subjects on the level of current events, allowing for occasional repetitions or paraphrased statements.

Level 3: Professional Working Proficiency

Speaking—Subject matter: The examinee is able to converse and express opinions about such topics as current events, including political and social problems of a general nature.

Speaking—Quality: The examinee has good control of grammar, though there are occasional errors in low-frequency structures and in the most complex frequent structure. The vocabulary is broad enough so that he or she rarely gropes for words in discussing the topics mentioned above. A foreign phonology, though apparent, is no longer dominant.

Understanding: The examinee can comprehend most of what is said at a normal conversational rate of speech. A person at this level is able to understand to a high degree more complex formal discourse, i.e., subjects on the level of panel discussion, new programs, etc.

Level 4: Distinguished Proficiency

Speaking—Subject matter: Although the subject matter that the examinee is able to handle at this level may not differ very much from that of Level 3, he or she is able to use the language in all nontechnical situations and can express opinions almost as fully and correctly as in native language (assuming that the individual is a 5 in the native language). The examinee is able to tailor his or her speech to the audience, has near-perfect grammar and speaks the language with extensive and precise vocabulary. Although the examinee may still have an accent, he or she very rarely mispronounces the language.

Understanding: The examinee can understand the content of all conversations and formal presentations within the range of his or her experience. With the exception of dialect variations and colloquialisms outside the range of experience, understands the type of language heard in speeches sprinkled with idioms and stylistic embellishments.

Level 5: Native or Bilingual Proficiency

BEST COPY AVAILABLE

Table 2. *Relationship between the ILR scale and the ACTFL scale (ACTFL, 1989, pp. 2-15)*

ACTFL SCALE	ILR SCALE	
	5	Native or bilingual proficiency
	4+	
	4	Distinguished proficiency
	3+	Superior
	3	Professional working proficiency
Advanced High	2+	
Advanced	2	Limited working proficiency
Intermediate High	1+	
Intermediate Mid	1	Survival proficiency
Intermediate Low		
Novice High	0+	
Novice Mid	0	No practical proficiency
Novice Low		

Table 3. *ACTFL Rating Scale (ETS, 1982, pp. i-iii)*

Novice Low (ILR'S Level 0)

*Unable to function in the spoken language.*Oral productions limited to occasional isolated words. Essentially no communication ability.

Novice Mid (ILR'S Level 0)

Able to operate only in a very limited capacity within very predictable areas of need. Vocabulary limited to that necessary to express simple elementary needs and basic courtesy formulae. Syntax is fragmented, inflections and word endings frequently omitted, confused or distorted and the majority of utterances consist of isolated words or short formulae. Utterances do not showevidence of creating with language or being able to cope with the simplest situations. They are marked by repetition of an interlocutor's words as well as by frequent long pauses. Pronunciation is frequently unintelligible and is strongly influenced by first language. Can be understood only with difficulty, even by person such as teachers who are used to speaking with non-native speakers.

Novice High (ILR'S Level 0+)

Able to satisfy immediate needs using learned utterances. There is no real autonomy of expression, although there are some emerging signs of spontaneity and flexibility. There is a slight increase in utterance length but frequent long pauses and repetition of interlocutor's words may still occur. Can ask questions or make statements with reasonable accuracy only where this involves short memorized utterances or formulae. Most utterances are telegraphic and word endings are often omitted, confused or distorted. Vocabulary is limited toareas of immediate survival needs. Can produce more phonemes but when they are combined in words or groups of words, errors are frequent and, in spite of repetition, may severely inhibit communication even with person used to dealing with such learners. Little development in stress and intonation is evident.

Intermediate Low (ILR'S Level 1)

Able to satisfy basic survival needs and minimum courtesy requirements. In areas of immediate need or on very familiar topics, can ask and answer simple questions, initiate and respond to simple statements, and maintain very simple face-to-face conversations. When asked to doso, is able to formulate some questions with limited constructions and much inaccuracy. Almost every utterance contains fractured syntax and other grammatical errors. Vocabulary inadequate to express anything but the most elementary needs. Strong interference from L1 occurs in articulation, stress and intonation. Misunderstandings frequently arise from limited vocabularyand grammar and erroneous phonology but, with repetition, can generally be understood by native speakers in regular contact with foreigners attempting to speak their language. Little precision in information conveyed owing to tentative state of grammatical development and little or no use of modifiers.

Intermediate Mid (ILR'S Level 1)

Able to satisfy some survival needs and some limited social demands. Some evidence of grammatical accuracy inbasic construction, e.g., subject-verb agreement, noun-adjective agreement, some notion of inflection. Vocabulary permits discussion of topics beyond basic survival needs, e.g., personal history, leisure time activities. Is able to formulate some questions when asked to do so.

Intermediate High (ILR'S Level 1+)

Able to satisfy most survival needs and limited social demands. Developing flexibility in a range of circumstances beyond immediate survival needs. Shows spontaneity in language production but fluency is very uneven. Can initiate and sustain a general conversation but has little understanding of the social conventions of conversation. The commoner tense forms occur but errors are frequent in formation and selection. Can use most question forms. While some word order is established, errors still occur in more complex patterns. Cannot sustain coherent structures in longer utterances orunfamiliar situations. Ability to describe and giveprecise information is limited. Aware of basic cohesive features (e.g., pronouns, verb inflections), but many are unreliable, especially if less immediate in reference.Extended discourse is largely a serious of short, discreteutterances. Articulation is comprehensible to nativespeakers used to dealing with foreigners, and can combine most phonemes with reasonable comprehensibility, but stillhas difficulty in producing certain sounds, in certainpositions, or in certain combinations, and speech willusually be labored. Still has to repeat utterances frequently to be understood by the general public. Able to produce narration in either past or future.

Advanced (ILR'S Level 2)

Able to satisfy routine social demands and limited work requirements. Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family, and autobiographical information; can handlelimited work requirements, needing help in handling any complications or difficulties. Has a speaking vocabulary sufficient to respond simply with some circumlocutions; accent, though often quite faulty, is intelligible; can usually handle elementary constructions quite accurately but does not have thorough or confident control of the grammar.

Advanced High (ILR'S Level 2+)

Able to satisfy most work requirements and show some ability to communicate on concrete topics relating to particular interests and special fields of competence. Often shows remarkable fluency and ease of speech, but under tension or pressure language may break down. Weaknesses or unevenness in one of the foregoing or in pronunciation result in occasional miscommunication. Areas of weakness range from simple construction such as plurals, articles, prepositions, and negatives to more complex structures such as tense usage, passive constructions, word order, and relative clauses. Normally controls general vocabulary with some groping foreveryday vocabulary still evident.

Superior (ILR'S Level 3 and Above)

All performance above Advanced High is rated as Superior.

Table 4. *Useful Question Types and Topic Areas*

Levels	Types of questions	Examples	Disadvantages	Topic Areas
0 - 0+ (ILR) Novice Low-High (ACTFL)	Yes/No questions	"Do you live in ...?" "Are you a student at the university?"	• Yes/No questions often provide no real information about the candidate's speech because they are so amenable to a one-word answer. • To encourage conversation, Yes/No questions must be followed by Wh-questions, such as "Who?", "What?", "Where?", and "When?"	Name articles of clothing Name basic objects Name colors Name family members Give weather Name weekdays, months Give today's date Give year, tell time
	Choice questions	"How did you get to work this morning, by bus or by car?" "Which do you like better, Japanese style breakfast or Western style breakfast?"	• Choice questions give away vocabulary and sometimes grammar points that the interviewer may be trying to check. Therefore, they may deprive the candidate of an opportunity to produce speech on her own.	
1 -1+ (ILR) Intermediate Low-High (ACTFL)	Polite request	"Would you describe this room, please?"	• Polite requests have to be carefully phrased in order to encourage speech production.	Personal information Hotel, restaurant Money matters Welfare Directions Transportation Meeting Social Telephone Post office Car
	Information questions	"Who was with you?" "When were you there?" "How did you get there?"	• The speech sample elicited by Information questions may be very short, in some instances only one or two words. • If candidates are not talking naturally in the target language, too many Information questions may lead them to feel that they are being interrogated. • Testers may end up rating the factual content of the speech sample, rather than its linguistic content.	
	Familiar situations (role-play)	"Please reserve a hotel room with a double bed and bath at the cheapest possible rate. I will play the role of the clerk."	• Acting out Familiar situations can be a problem to any candidate who dislikes role-playing.	
	Candidate interviews Testers	"Please ask me some questions, such as where I live, etc."	• The candidate may simply repeat the questions which the tester gives as illustrations. • Candidates often draw a blank when asked to produce speech out of context, such as "I've asked a lot of questions. Now I'd like you to ask me some questions."	
2 -2+ (ILR) Advanced & Advanced High (ACTFL)	Information questions		(see above)	Like the Level 1 speaker, the linguistic repertoire is generally in the realm of "who," "what," "where," and "when," rather than "how" and "why" (a distinctive feature of the Level 3 speaker).
	Familiar situations with complications (role-play)	"You are in a restaurant. You have eaten most of your meal when you discovered a bug under the steak. You feel ill. Call the waiter and announce that you are leaving immediately. Refuse to pay the bill." "You live in a condominium. Your upstairs neighbor waters the plants on her balcony and the water ends up on your balcony, dirtying the washings. Go to your neighbor and discuss the problem. I will play the part of the neighbor."	• • Some candiates dislike role-playing. • Role-plays may turn into a translation exercise, if the situations are too specific.	Ask others for information about themselves, such as family status, residence, background and interests. Give this type of detailed information about herself. Describe daily routine and talk about one's personal interest. Describe a person, an object or a place. Give directions. Tell about a sequence of events. Tell in simple terms about plots of books or movies (Detailed discussion of them is a Level 3 task). Tell about news events in simple factual terms. Tell about future plans.

(based on ETS, 1982, pp. 43-57, 75-86)

BEST COPY AVAILABLE

Table 5. *Simplified Checklist of ILR/ACTFL Level Descriptions*

0+ / Novice High	1 / Intermediate Low & Mid	2 / Advanced	3 / Superior
1) Very little or no ability to create future • One-to two-word responses • Frequent repetition of interviewer's word through Q & A • Very limited communication using formulas, memorized phrases) 2) Interviewer must repeat often	1) No control of past and future (most speech in present tense) 2) Can create original utterances 3) Can communicate simply • Can make simple descriptions about personal background • Can ask simple questions, directions and instructions • Can handle minimum courtesy requirements • Areas of immediate need, or very familiar topics 4) Almost all utterances contain grammatical errors	1) Can sustain control in the past, present, and 2) Can narrate, and describe 3) Real converstion (not just Q & A) 4) Can provide limited discourse 5) Can get out of survival situations with complications	1) Can discuss hypothetical situations 2) Can support opinions 3) Can provide extended discourse 4) Good control of grammar; only occasional errors 5) Has broad vocabulary and rarely gropes for words 6) Can handle unfamiliar topics or situations

(Pluses (+) indicate ability at almost the next level, not halfway ranges)

1+ / Intermediate High
- Can narrate and describe but unable to sustain control in past and future

2+ / Advanced High
- Supports opinion and hypothesizes in a limited manner.

Chapter 13

The SPEAK Test of Oral Proficiency: A Case Study of Incoming Freshmen[1]

SHAWN M. CLANKIE
KANSAI GAIDAI UNIVERSITY

The SPEAK test, or *Speaking Proficiency English Assessment Kit,* was created by the Educational Testing Service (for more information, see ETS, 1993), the makers of other tests including TOEFL and TOEIC, and was developed "in response to the interest expressed by many institutions in an instrument to assess the spoken proficiency of foreign teaching assistants and other international students who are non-native speakers of English" (ETS, 1992, p. 5). This chapter presents the results of an attempt to adapt the SPEAK test into an existing English as a foreign language program at Kansai Gaidai University in Osaka. It will examine in depth the benefits of the SPEAK test, some of its problems, and the reasons why we chose to abandon the test, opting for a different form of oral evaluation.

The Program

In the spring of 1993, Kansai Gaidai University instituted a new program called the IES, or Intensive English Studies program. This program coincided with the existing, less demanding program, and English majors entering the university were allowed the option of automatically entering the regular English program or trying to test into the IES program. The program was inaugurated at both the junior college and university levels, with three native speaking university level instructors and two native speaking junior college instructors.

For those who applied for the intensive courses, the IES program was designed to afford the top 10-15% of the students the opportunity to study the language intensively. This program offered approximately double the standard number of class contact hours per class per week with a restricted number of different native speaker instructors. Class size was restricted as well, to a maximum of 30 students per class. Classes were taught completely in English, and it was hoped that the increased exposure to English with a limited number of good students and teachers would foster higher levels of proficiency than those reached in the existing program, particularly in speaking ability at the end of the respective four or two year terms.

The number of students in the initial year was 85 and 47 at the university and junior colleges, respectively. The 132 charter students were selected from a group of several hundred on the basis of TOEFL scores alone and were placed by those same TOEFL scores into an IES section of students with similar scores. The university level consisted of three sections (A, B, and C), while the junior college had two sections (A and B).

Selection by TOEFL scores alone posed a potential problem in that the students selected, while strong in grammar and writing skills were not necessarily the strongest in terms of oral production. This brought up the possibility that the program might be losing students who have strong oral skills but who lack the skills necessary for success on the TOEFL. Realising this problem, the IES teachers began seeking alternatives to selection strictly on the basis of TOEFL, in particular looking for a measure to assess oral ability, that could be administered quickly and used along with the TOEFL scores to gain a truer picture of the students with the best ability.

In looking for a test of oral ability, we had several concerns. We first needed a measure that could be used to test a large number of students. We were worried that it might be impossible to accurately assess several hundred students one at a time given scheduling and staff limitations. Time was a second major factor. The school year begins on April 8, and, from the time of acceptance into Kansai Gaidai University, the students would need to apply to the IES program, be TOEFL tested, be selected into the IES program, be placed into a section of the program, and be notified of acceptance and scheduling, all in a matter of a couple of weeks. The administration of an oral measure would of course be in addition to the procedural steps above.

The SPEAK Test

In November of 1993, after a semester of looking at alternatives to selection strictly on the basis of TOEFL scores (and ever mindful of the time constraints mentioned above), the teachers of the IES program at Kansai Gaidai University began attempts to integrate the SPEAK into a workable model for assessing the skills of incoming freshman seeking entrance into the IES program in the spring of 1994. In addition to this, an attempt was made to find a way to combine SPEAK scores with the TOEFL scores as the basis for admissions and placement decisions in the IES program.

The first trials of the SPEAK test were administered in November and December of 1993 to the three charter classes of university level second semester freshman and the two charter classes of junior college second semester freshman already enrolled in the IES program. SPEAK was selected because it could easily be administered to large groups with fairly high interrater reliability.

The test kit comes with one set of thirty test booklets, the test cassette, and a guide to administering the test, as well as scoring sheets, rater training cassettes, and instructions. Additional versions of the SPEAK test are also available. It should be noted here that the SPEAK test is generated from the TSE (Test of Spoken English). Retired TSE tests make up the three versions of SPEAK currently on the market. The cost of the kit at present is $350 with additional versions of the test for $150 each. All of the tests are of course reusable and would work best in a rotation system wherein each version of the test is used once in three years or semesters.

The actual testing procedure simply involves distribution of the test booklets and the playing of the matching test cassette. Students respond to each question or problem and their voices are recorded on separate cassette recorders.

The test contains seven sections, of which the last six are scored. The first section is to relax the students and to test the equipment and allow for modifications with questions such as "What is your name?" In section two, students are asked to read a paragraph silently to themselves, then when instructed, to read the paragraph aloud. In section three, students are given the beginnings of sentences and are asked to complete them. This is followed by a picture sequence used to elicit a narrative in section four. Section five involves a single picture from which questions are asked to the students about what is, will, or has happened. Section six contains open-ended description questions, and students are asked to explain. The final exercise asks the students to present a class schedule as if they were a teacher.

For each of the six scored sections of the test, students are scored in at least two of four categories. The categories are pronunciation, fluency, grammar, and comprehension. Each answer is rated on a scale of 0-3, three being the best. At the end of the test, the scores of each of the core categories are added together then averaged, with the exception of the comprehension segment where the average is multiplied by 100. Then scores are compared and matched to a chart defining the abilities of the student.

Raters are trained according to several sample test cassettes that come with the test package. The students' cassettes then are distributed among the raters, who rate them, then meet to discuss and select students to enter the program. This all seemed easy enough.

Our purpose for testing the students already taking part in the program was primarily to gain experience with the test and to work out any glitches found before actually putting the test into practice. This first run was important as we quickly found problems in using the SPEAK to assess Japanese learners. As the students taking part in the experiment were already in sections, we tested each section one class at a time over the span of a week. The tapes recorded by each student were then collected and each was rated by two instructors (in the hopes of increasing reliability and validity). In the case of the two junior college instructors (of which I was one), this meant each of the two of us took one of the two sections of junior college classes' tapes, rated them, then exchanged the tapes for the other section and rated those as well. The results of the two raters were then matched to see if the scores were similar, both in comparing students and classes.

Benefits of SPEAK

The obvious benefit of the SPEAK test is that it can be administered to a group or a single student with the same ease. Moreover, the problem of some interviewers being more difficult in questioning or scoring than others is overcome in that all students take the same test, administered the same way, with the students each having the same amount of time to respond. And of course, there is little thinking time, students are expected to respond immediately, as in normal conversation. This prevents (in most cases) those long thinking pauses which often stall face-to-face oral interviews. There is another plus. If several versions of the test are at the disposal of a program, they may be rotated so that each entering group of students takes a different version of the test.

Sections of the test such as those which include free description of an open question or topic reflect tasks which are a definite practical concern in daily conversation. This test appears ideal for cases where one student needs to be interviewed, i.e., a graduate student seeking a teaching assistantship.

One final benefit, the test saves the time that it takes to create the lists of questions and topics which are normally used in face-to-face interviews. In the case of a small program such as ours, it would appear that such benefits would make the test very useful. Yet, we found significant drawbacks to using this test in its published format.

Drawbacks of SPEAK

The primary fault of the test (for our purposes) is generated by one of its most outstanding benefits, i.e., the fact that it can be given to a large group at the same time. The traditional oral examination is face-to-face with a live tester who is administering the test and rating the examinee during or directly after the interview. This means that each examination is given once, and once it's finished it's finished.

However, the SPEAK test is the same test every time on every cassette, with varying levels of differences in responses. With all of the students taking the test at the same time and their answers being recorded, each tape must be listened to and rated. In essence, what one gains in giving the test quickly is lost by the amount of time it takes to listen to all of the cassettes. With each test being roughly twenty minutes long, multiplied by a class of 25, a rater continuing non-stop must work over 8.5 hours. This of course is nearly impossible, and we found that a severe drop in rater reliability as well as an increase in daydreaming during the rating, and in particular, during the lull periods between sections, began to take effect after only 3-5 tapes.

As mentioned in the opening of this essay, time was critical, as all rating, selection, and notification had to be done prior to the beginning of the semester. In addition to the students' voices, the original taped examiners' voices are also included on the tapes. Listening to the same test repeatedly had the effect of distracting the raters from concentrating on the intermittent test segments containing students' responses. Trying to fast forward only to the answers was a hit or miss solution at best and, in any case, did not significantly reduce listening time.

Rater willingness to listen to the cassettes after the initial day of administration was another problem. Enthusiasm was high at the onset of the experiment. However, after the first day of listening to cassettes, the raters found themselves increasingly unwilling to continue with the SPEAK test ratings.

In addition, one particular part of the test, the second exercise, which asks students to read aloud is somewhat questionable in terms of test validity. Hughes (1991, p. 110) discourages the use of reading aloud as a measure of oral proficiency because of "inevitable interference between the reading and speaking skills." He also argues that, if the same test were given to a group of native speakers, there would "almost certainly be considerable differences between candidates." Underhill (1987, p. 76) also supports this view.

Moreover, we found that the scoring system needed some modification to bring it more into line with the likely responses that we anticipated receiving. While the test would seemingly work well for students possessing widely varying levels of mastery in English, it simply was not satisfactory when given to large numbers of students of similar abilities. Needless to say, it is well known that Japanese students leaving high school have years of training in areas such as grammar, yet in most cases, their oral abilities are weak. The scoring system accompanying the test, based on a 0-3 rating simply could not differentiate in its mid-range of 1-2. Obviously, it was easy to spot students at the high end, often those with substantial overseas experience for example, yet in the mid-range the task was much more difficult.

A similar problem arose when we discovered that some of the responses given by our students were direct exceptions to the scoring scale created by ETS. One question offered by one of the raters in our program involved how to score the student who uses a longer, more complex utterance containing several errors versus the student who only answers in short yet perfectly correct responses. The student offering the longer responses was apparently trying to do his or her best while the other student was simply taking the safe way out.

In the end, after attempting to rate a portion of the cassettes according to the prearranged rating scale, we found that change was needed. A meeting was held and each of the teachers involved presented suggestions for making the scale more effective, primarily in the mid-levels. After sifting through each of the four rating categories (pronunciation, fluency, grammar, and comprehensibility), a certain level of success in adding additional half-levels in the mid-ranges seemed to offer greater differentiation in scoring. The original and adapted scoring scales are presented in Tables 1 and 2, respectively (see the Appendix to this chapter).

Hughes (1991, p. 105) points out a further disadvantage of group testing from a cassette when he states, "The obvious disadvantage to this format is its inflexibility: there is no way of following up candidates' responses."

As a final note, on taped tests such as the SPEAK, Underhill (1987) also mentions simple practical problems such as the possibility of technical difficulties which could arise causing some or none of the tapes to be recorded. This actually happened on our first attempt to administer the test to a class and is definitely a practical concern.

Prospects for SPEAK and Conclusions

The question now arises as to what (if anything) can be done to make the SPEAK test applicable to the various English programs existing in Japan. If the test is simply to be used to check a handful of students' abilities, the test will adequately serve this purpose. Obviously, the more raters that can be involved in the test the better the chance for success. The less tapes corrected by each rater, the more reliable will be the ratings. Reducing the number of tapes rated at a single sitting to three will likely also increase rater reliability, but will also take substantially more time to rate. With new teachers being added each semester in our particular program, the possibility of reducing the number of tapes listened to by each rater has increased. As of summer 1994, 14 teachers were involved in the IES junior college and university programs. However, with an incoming group of potential IES students of 400 (after some form of preliminary elimination by TOEFL), that would mean that each teacher would be responsible for listening to 28.5 tapes or slightly more than a class. That is only if the tapes are to be listened to by a single rater. With each tape being 20 minutes, the time per rater amounts to 9.5 hours. This is still far too high.

Moreover, repeatedly listening to the original test prompts on each of the students' cassettes is still a problem. For schools that contain language labs with voice activated recorders this problem could be overcome by taping only the voice of the students and thereby eliminating several minutes of blank tape waiting time. As many schools are not yet equipped with such a system, this problem may still plague users of the SPEAK test. Finally, some schools may find the test more successful if they manipulate the scoring systems to more accurately represent the problems and abilities of the Japanese learners in their individual programs (as we did in Table 2). Inevitably, some schools trying out the SPEAK test may choose to abandon the test altogether, as was the case at Kansai Gaidai.

At Kansai Gaidai, given the constraints of our system, we found that the TOEFL scores in conjunction with an oral interview, one consisting of two interviewers interviewing each student one at a time was a more effective and time efficient option than the SPEAK test. This has allowed us to interview, rate, and place the students all in the course of a single day.

We held a faculty meeting with the administration and the department coordinators and arranged a testing day, much the same way as when we were to conduct the SPEAK test. This was held roughly a month ahead of the actual oral testing date, allowing ample time for logistical concerns (scheduling interview rooms, notifying candidates, etc.).

Then the teachers met a couple of days prior to the test (e.g., over one or two lunch hours) to discuss possible topics for the interview. These were selected from current events, both international and domestic, and from general experiences such as overseas travel, and from general conversation topics. On the testing day, each student was then interviewed by two teachers in the program, in roughly an eight minute interview, then the scores were compared to the TOEFL scores. It was possible to have seven interviews occurring simultaneously throughout a three hour morning session. The morning session consisted of interviewing all of the junior college candidates.

Scores were collected hourly by officials of the registrars' office and entered into the computer. Over a working lunch period the faculty selected the students permitted to enter the two IES junior college classes. In the afternoon session the university candidates were rated in the same process. Upon completion of the afternoon session, we examined the results and selected those who would take part in the four university-level IES classes. At the end of the one day testing session, the teachers were taken to dinner in gratitude for the day's work (an important factor in assuring teachers to volunteer to take part the following term). This manner of testing, of course, has taken away some of the objectivity which the SPEAK test offers in that the same test is not being offered to each student.

As time is among the most pressing concerns for university English programs in Japan, the form of test used, its administration, rating, and the reason for testing must all be carefully scrutinised. For those considering using the SPEAK test, this chapter should present a clearer picture of how the test may work in assessing students in a Japanese academic situation.

Note

[1] I would like to give special thanks to Mary Everett at Kansai Gaidai for preserving many of our notes regarding this experiment, and for reading an earlier draft of this essay.

References

ETS (Educational Testing Service). (1992). *Guide to SPEAK*. Princeton, NJ: Educational Testing Service.

ETS. (1993). *TOEFL 1993-1994 products and services catalogue*. Princeton, NJ: Educational Testing Service.

Hughes, A. (1991). *Testing for language teachers*. Cambridge: Cambridge University Press.

Underhill, N. (1987). *Testing spoken language*. Cambridge: Cambridge University Press.

Appendix: Tables

Table 1. *SPEAK Scoring Key*[*]

Pronunciation	
0	Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be unintelligible.
1	Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be occasionally unintelligible.
2	Some consistent phonemic errors and foreign stress and intonation patterns, but speaker is intelligible.
3	Occasional non-native pronunciation errors, but speaker is always intelligible.
Fluency	
0	Speech is so halting and fragmentary or has such a non-native flow that intelligibility is virtually impossible.
1	Numerous non-native pauses and/or a non-native flow that interferes with intelligibility.
2	Some non-native pauses but with a more nearly native flow so that the pauses do not interfere with intelligibility.
3	Speech is as smooth and as effortless as that of a native speaker.
Grammar	
0	Virtually no grammatical or syntactical control except in simple stock phrases.
1	Some control of basic grammatical constructions but with major and/or repeated errors that interfere with intelligibility.
2	Generally good control in all constructions with grammatical errors that do not interfere with overall intelligibility.
3	Sporadic minor grammatical errors that could be made inadvertently by native speakers.
Comprehensibility	
0	Overall comprehensibility too low in even the simplest type of speech.
1	Generally not comprehensible due to frequent pauses and/or rephrasing, pronunciation errors, limited grasp of vocabulary, or lack of grammatical control.
2	Comprehensible with errors in pronunciation, grammar, choice of vocabulary items, or infrequent pauses or rephrasing.
3	Completely comprehensible in normal speech with occasional grammatical or pronunciation errors.

[*] Reprinted by permission of Educational Testing Service, the copyright owner (ETS, 1992, p. 10). No endorsement of this publication by Educational Testing Service should be inferred.

BEST COPY AVAILABLE

Table 2. *Revised Scoring for SPEAK, Kansai Gaidai University*

Comprehensibility	
0	Does not comprehend the question.
1	Generally not comprehensible; with pauses and errors; limited vocabulary.
1.5	Overall comprehensibility low; many errors.
2	Generally comprehensible; short answers.
2.5	Completely comprehensible in normal speech with occasional grammatical or pronunciation errors.
3	Easily comprehensible with lengthy answers.
Fluency	
0	Too halting or fragmentary to be intelligible.
1	Numerous non-native pauses and/or a non-native flow that interferes with intelligibility.
1.5	Numerous pauses but fairly intelligible.
2	Some non-native pauses but with a more nearly native flow so that the pauses do not interfere with intelligibility.
2.5	Pauses but intelligible; lengthy answers.
3	Smooth, effortless speech with few errors.
Grammar	
0	Virtually no grammatical or syntactical control except in simple stock phrases.
1	Some control of basic grammatical constructions but with major and/or repeated errors that interfere with intelligibility.
1.5	Good control of basic constructions; fair intelligibility.
2	Generally good control in all constructions with grammaticalerrors that do not interfere with overall intelligibility.
2.5	Good control; good intelligibility; more complex structures.
3	Sporadic minor grammatical errors that could be made inadvertently by native speakers.
Pronunciation	
0	Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be unintelligible.
1	Frequent errors; often not intelligible.
1.5	Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be occasionally unintelligible.
2	Some consistent phonemic errors and foreign stress andintonation patterns but speaker is intelligible.
2.5	Some errors but speaker is generally intelligible.
3	Occasional errors but speaker is always intelligible.

Chapter 14

Making Speaking Tests Valid: Practical Considerations in a Classroom Setting

Yuji Nakamura
Tokyo Keizai University

In this chapter, I will (a) examine various kinds of validity in language testing in general, (b) discuss which types of validity are most suitable for a test of English speaking ability in a classroom setting, and (c) describe a validity case study which illustrates the types of validity (i.e., construct validity, concurrent validity, face validity, and washback validity) which can be used in the development of a semi-direct speaking test suitable for Japanese college students.

Various Kinds of Validity in Language Testing

Spolsky (1975) says that the central problem in foreign language testing, as in all testing, is validity. Messick (1988) further states that although the modes and methods of measurement may change, the basic maxims of measurement, and especially of validity, will likely retain their essential character. Although there are other aspects of testing (such as reliability and practicality) which are crucial and should not be overlooked in language testing, there is no doubt that validity is the single most critical element in constructing foreign language tests.

Validity concerns the question of "How much of an individual's test performance is due to the language abilities we want to measure?" (Bachman, 1990, p. 161). In other words, it is concerned with how well a test measures what it is supposed to measure (Thrasher, 1984).

Traditionally, validity has been classified into several types including at least content, criterion-related, and construct validity (Bachman, 1990). These are all concerned with the relationship between the test and the domain to be measured (Davies, 1990). Further, Messick (1993) and APA (1985) argue that validity as a unitary concept.

Davies (1990) discusses five kinds of validity (face, content, construct, predictive, and concurrent). Among these, predictive validity and concurrent validity are usually called criterion-related validity or external validity because they have criteria outside of the proposed new test. Content validity and construct validity, on the other hand, are concerned with internal aspects of validity.

Thrasher (1984) classifies these five types of validity into two categories. One includes predictive, concurrent, and construct validity, the other face and content validity. He claims that the former three are prestigious because elegant statistical processes are available for computing them, while the other two lack prestige because no such statistical processes can be applied to them.

In the following discussion, I will follow Bachman's (1990) view that we still find it necessary to gather information about content validity, predictive validity, concurrent validity, etc., and then investigate validity further from more detailed points of view. He also mentions the following issues as being relevant to validity: (a) cultural background, (b) background knowledge, (c) cognitive characters, (d) native language, (e) sex, (f) age, (g) ethnicity, and (h) washback/backwash effect (or the effect of testing on curriculum).

Interestingly, Thrasher (1984) adds a new type of validity which he calls educational validity. This type of validity applies to the relationship between positive test effects and students' study habits, among other variables, by taking into consideration the educational and testing situation in Japan. This is similar to Bachman's claim that positive washback will result when the testing procedures reflect the skills and abilities that are taught in the courses.

In this chapter, I will discuss the various types of validity, basically following Thrasher's (1984) categorization of validity in the following order: predictive, concurrent, construct, face, content, and educational validity.

Predictive Validity

Thrasher (1984) claims that predictive validity is an estimate of the goodness-of-fit between test results and performance in some posttest real world activity. Consequently, predictive validity is concerned with the predictive force of a test, the extent to which test results predict some future outcome (Davies, 1990). As we know from the TOEFL (*Test of English as a Foreign Language*) results, the ideal of test results and performance in the posttest real world activity is very important in the predictive validity. This is an appealing aspect of predictive validity and the associated statistics, since prediction is an important and justifiable use of language tests.

Predictive validity, however, also has some weak points. Since this type of validity compares a test with some criterion that is measured after the test results are in, the nature of the criterion skill or ability becomes very important. TOEFL results, for example, have been compared to college and graduate school performance, yet such performance is obviously affected by many factors other than English proficiency. In other words, measures that are valid predictors of some future performance are not necessarily valid indicators of language ability (Bachman, 1990). This point is crucial because language tests should measure language abilities, and nothing else. Therefore, we should be careful deciding what abilities are being measured.

Concurrent Validity

Davies (1990) says that concurrent validity, in its purest form, can be established only when the test under scrutiny represents either a parallel or a simplified version of the criterion test, and, as Hughes (1989) says, when the test and the criterion are administered at about the same time. This last condition is one of the things that differentiates concurrent validity from predictive validity as types of criterion-related validity.

Construct Validity

Bachman (1990) states that construct validity concerns the extent to which performance on a test is consistent with predictions that we make on the basis of a theory of abilities or constructs. In brief, construct validity examines if the test matches a theoretical construct (Thrasher, 1984).

Bachman further insists that construct validity is a unifying concept, which is supported by Messick's (1993) idea that construct validity is the unifying concept that integrates criterion and content considerations. What is it that constitutes this unifying concept of construct validity?

Bachman (1990) claims that construct validity requires both logical analysis and empirical investigation. Logical analysis is involved in defining the constructs theoretically and operationally. Construct validity has both strong points and weak points. One strength is that it can be statistically analyzed through the multitrait-multimethod approach or factor analysis, which avoids the reduction problem due to the external criterion which occurs in concurrent validity. One weak point is that testers tend to view technical terms (such as grammar, vocabulary, reading) as almighty words which are representatives of theoretical constructs and can measure the whole range of language abilities. Since the word *construct* refers to any underlying ability (or trait) which is hypothesized in a theory of language ability, and since this construct is complex, construct validity cannot be established overnight.

Moreover, as Hughes (1989) says, it is through construct validation that language testing can be put on a sounder, more scientific footing, and this fact is the strongest argument for establishing construct validity.

Face Validity

Thrasher (1984) questions the utility of face validity when he asks "what is the meaning of the word 'reasonable' in the sentence 'the test looks reasonable.'" However, I support Heaton's (1989) belief that "in the past, face validity was regarded by many test writers simply as a public relation exercise. Today, most designers of communicative tests regard face validity as the most important of all types of test validity."

Face validity is hardly a scientific concept, but it is very important (Hughes, 1989). Test appearance (face validity) is a very important consideration in test use because a test which does not have face validity may not be accepted by candidates, teachers, etc. (Bachman, 1990). The most important part of face validity in school tests is that the students' motivation is maintained if a test has good face validity, for most students will try harder if the test appears fair (Heaton, 1989). The weakest aspect of face validity is also discussed by Heaton (1989) when he says that language tests which have been designed primarily for one country and are adopted by another country may lack face validity in the new setting. In general, tests usually have high face validity because they look like other tests of that skill that the student has previously taken. So if a test from one country does not look like the tests with which students are familiar in their own country, it will be weak in face validity.

Finally, face validity cannot stand alone. More precisely, in order to make face validity convincing, we must have very strong evidence of the other types of validity (cf. Thrasher 1984).

Content Validity

Thrasher's (1984) statement about content validity is interesting. He says that even if the language teacher does not consider content validity, his/her students will demonstrate the need for it (probably by analyzing, criticizing, or complaining about the content of the test).

As Davies (1990) says, content validity is defended through professional judgments, either by teachers or testers. The judges rely on their knowledge of the language to judge to what extent the test provides a satisfactory sample of the syllabus, whether real, imagined, or theoretical.

A test is said to have high content validity if its components constitute a representative sample of the language skills and structures that it is meant to test (Hughes, 1989). Accordingly, a careful analysis of the language being tested and of the particular course objectives is needed to demonstrate content validity. Two possible strong aspects of content validity are as follows:

1. We must consider a wide variety of knowledge and content coverage (e.g., definition of the content, ability domain, a list of content areas, test tasks, etc.) (Bachman, 1990).
2. The greater a test's content validity, the more likely it is to be an accurate measure of what it is supposed to measure (Hughes, 1989).

However, there are some weak points, too. Firstly, content validity says nothing about the appropriateness of what has been taught; in other words, we cannot be sure if we are teaching the right thing (Thrasher, 1984). Secondly, content validity (content relevance in Bachman, 1990) by itself is inadequate as a basis for making inferences about abilities because content validity looks only at the test and does not consider how test takers perform (Bachman, 1990). Thirdly, if content validity is misunderstood and the content of a test is determined by what is easy to test rather than what is important to test, the test is likely to have a harmful washback effect. Areas which are not tested are likely to become areas ignored in teaching and learning (Hughes, 1989).

Since content validity has been the primary type used in achievement testing—the kind of measurement language teachers must do in their classrooms (Thrasher, 1984), the weak points related to the appropriateness of teaching should be reconsidered.

Educational Validity

As mentioned above, Thrasher (1984) offers educational validity as a sixth validity in addition to the other five standard validities (predictive, concurrent, construct, face, and content). After pointing out the characteristics of these five validities, he was still not satisfied because of the weak points of content validity, especially in terms of classroom tests (for achievement) rather than qualification tests (for proficiency). His reason for feeling this way was explained by one statement: ". . . content validity says nothing about the appropriateness of what has been taught although it tells you if the course and test content match," and, ". . . we are not sure if we are teaching the right thing."

This is closely related to his criticism of Spolsky's (1975) statement "foreign language tests used by classroom teachers have few problems in validity, because the textbook or syllabus writer has already specified what should be tested."

Thrasher (1984) thought that content validity was not sufficient from the viewpoint of the appropriateness of teaching, and he suggested using the notion of educational validity for considering the tight relationships among testing, teaching, study habits, test results, and course objectives in terms of the positive washback effects of the tests.

Although Bachman (1990) deals with washback effects regarding validity, Thrasher's (1984) idea of educational validity is more practical and relevant for Japan's educational system. Thrasher (1984) takes up the fact that college entrance exams have great influence in Japan. These tests have a crucial impact on what students learn, and he wonders if we can measure the result of the impact. Thrasher's concept of educational validity includes two assumptions: (a) any test has effects—good and bad—on student morale, study habits, and understanding of what the course of study is trying to accomplish, (b) such effects can be measured and judged beneficial or detrimental to the goal the teacher has laid down (Thrasher, 1984). He assumed two hypothetical objections: (a) test effects cannot be measured, and (b) study habits are determined not only by the test but also by a complex of cultural, psychological, and educational background factors. Thrasher refuted both objections convincingly, in my opinion.

Coincidentally or not, Bachman (1990) points out, in his definition of validity, some additional elements (such as the student's culture, educational background, washback effects, ethnicity, etc.) in addition to other traditionally established aspects of validity (such as logical or empirical analysis).

If educational validity can be established as a separate type of validity, I think classroom teachers will benefit greatly.

Types of Validity Suitable for Testing Speaking Ability

A test of speaking ability in a classroom setting is usually used as an achievement test. According to Davies (1990), an achievement test should have both face and content validities. I would argue that predictive validity and educational validity as well as construct and concurrent validities should also be analyzed.

First, as Davies says, content validity is unavoidable for a classroom speaking test which has the characteristics of an achievement test. Since content validity simply asks if the test content (vocabulary, grammar, and tasks) matches the content of the course of study, what testers (teachers) can do is to match the course objectives and syllabus design (which themselves should be based on construct validity) with the test items. This effort should reduce students' complaints about the content of the speaking test. In the traditional understanding of content validity, tasks are less important than the match between test and classroom vocabulary and grammar. This attitude toward tasks by teachers is crucial in a classroom test because teachers my unconsciously tend to use test tasks which are different from the course objectives especially when oral/aural aspects are involved.

Second, construct validity is concerned with matching the theory of speaking and the tasks the test-maker requires the test takers to perform. This cannot easily be handled by classroom teachers because of the technical and statistical analyses involved and because of the abstract nature of language abilities. Construct validity is the most fundamental validity for a speaking test, however, even if it is difficult to carry out, because the test tasks themselves (the speaking tasks in this case) are of primary concern in construct validity.

Third, face validity is a must in a classroom speaking test. Semi-direct speaking tests like tape-recorded tests have much more face validity than indirect tests of speaking skills like paper-and-pencil tests; accordingly, students' motivation is promoted and maintained for speaking, and test results will be more reliable. Although there is an argument that the move from multiple-choice to productive tests usually means reduced reliability, we need to make a distinction between task reliability and scorer reliability when it comes to speaking tests. Task reliability will be enhanced by having the students do something they believe is a valid speaking activity, which means focusing on task reliability. Students' speaking abilities should be measured in a test in which they think they are taking a speaking test. Ideally, a direct speaking test such as an interview test is the best; few institutions can, however, conduct interviews because of financial and practical considerations.

Fourth, predictive validity is feasible in many schools for the following reason: We can check students' posttest real world oral activities in English-speaking countries when they go overseas during the vacation, because many schools have a SEA (Study English Abroad) program and send students to English-speaking countries every year, for a few weeks to a year. We can compare students' in-class test results with their results in real world communication in English-speaking countries.

Next is educational validity. Although this is not an established form of validity, the idea of the relationship among testing, teaching, study habits, and test results from the viewpoint of positive washback effect is highly recommended in a classroom speaking test. If students change their study habits (from focusing on grammar-centered study to focusing on spoken English, listening to English, and trying to speak out), that is one of the objectives of the course and the speaking test. This is a case where a speaking test can have good effects on students. Although there may be a long way to go before this sort of validity can be established, educational validity may turn out to be an important factor in justifying the usefulness of a classroom speaking test to promote the speaking ability of Japanese students.

Lastly, concurrent validity may not be easy to determine if we only think in terms of having students take two tests at about the same time and comparing the results. Regrettably, we cannot find a relevant criterion test, since it should itself be valid, reliable, and practical. There are, of course, some imported speaking tests such as the FSI (Foreign Service Institute) Interview test or the ACTFL (American Council on the Teaching of Foreign Languages) OPI (Oral Proficiency Interview) test, for intermediate and high level students. However, those tests are not relevant for lower level students because of the difficulty level of the test items.

Rather than waiting for the completion of another speaking test, classroom teachers could use native speaker teachers' assessments as a criterion. We could use conversation teachers' class grades or their estimates of speaking ability as criteria and investigate concurrent validity by comparing the test results and teachers' class grades or estimates of speaking ability.

What we need to do for this purpose is to achieve high interrater reliability for these assessments. One way to do so is to find a reliable teacher as an estimator. Another is to conduct a comprehensive training session to create high interrater reliability among native speakers (NS). Still another way is to have NS estimators focus only on the evaluation of speaking ability rather than non-language aspects like attendance, effort, or submission of homework. Since there are many native speakers coming to secondary schools through the JET (Japan Exchange and Teaching) or AET (Assistant English Teachers) programs, we could measure students' oral activity outside the testing situation with the help of these native speakers.

Even in colleges, the number of NSs is increasing. We could compare students' in-class speaking test results with their performance in talking with NSs under quasi-real-world situations. This type of validity investigation might be more practical in a school situation because non-native English teachers can easily compare the students' test results in their language classes with the same students' actual behavior or performance in English in native English speaking teachers' classes.

A Case Study

Background of the Test Development

Performance testing (especially testing oral proficiency) has become one of the most important issues in language testing since the role of speaking ability has become more central in language teaching with the advent of the Communicative Approach. There is a great discrepancy, however, between the expansion of the communication boom and the accurate measurement of communication ability (especially speaking ability) because of the many difficulties involved in the construction and administration of any speaking test.

In this case study (Nakamura, 1993), I constructed a speaking test based on Bachman's Communicative Language Ability model (1990) and examined the detailed components of Japanese students' English speaking ability based on that model. I also checked the validity, reliability, and practicality of these new testing procedures.

Four Kinds of Validity Actually Used

The following four sub-categorizations of validity were examined in the this case study:

1. **construct validity**—whether the test matches a theoretical construct (Thrasher, 1984)
2. **concurrent validity**—whether a new test is measuring the same thing as another more established one (Thrasher, 1984)
3. **face validity**—whether the test looks valid or reasonable to the examinees who take it (cf. Weir, 1990)
4. **washback validity**—the effect of testing on teaching and learning (Hughes, 1989)

Other validities such as predictive validity or educational validity (Thrasher, 1984) were not dealt with individually in the present research. Predictive validity could not be implemented because of practical limitations, and educational validity is included in the broader definition of washback validity that I used.

Methods

Eighty college students took the test consisting of four tasks (Task I - Speech Making; Task II - Visual-Material Description; Task III - Conversational Response Activities; Task IV - Sociolinguistic Competence Test named Mini Contexts).

Eleven raters[1] (four Japanese and seven native English speakers), all of whom had been teaching English for at least one year, evaluated eighty audio tapes on which the students' responses had been recorded in the language laboratory. The raters used the scoring sheet and scoring criteria designed by the present author. The first two tasks were rated on a four point scale (below average, average, above average, very good) in each linguistic component (grammar, vocabulary, pronunciation, etc.). Conversational responses were rated on a different four point scale (no answer, conversationally inappropriate, conversationally appropriate, very good). Sociolinguistic competence answers were rated on yet a third four point scale (no answer, sociolinguistically inappropriate, sociolinguistically appropriate, very good).

Data Analysis

Each rater's raw score for each task was summed up to obtain a total score. Then, interrater reliability was measured through each rater's total score on the 80 tapes using Pearson's formula. The internal consistency was examined through Cronbach's alpha to establish another measure of reliability.

Four tasks were examined from the viewpoint of the content validity. The four tasks should be mutually exclusive and only moderately inter-dependent to be composites of the proposed framework of speaking ability.

Correlation coefficients were also calculated between the four tasks and independent criterion measures. The concurrent validity was examined by looking at the correlation between four tasks and teachers' class grades and a teacher's estimates. Factor analysis was conducted to examine the construct validity.

In the questionnaire analysis, I asked the 80 students who took the speaking test to answer a questionnaire on their impressions of the test from the viewpoint of their study habits toward the improvement of their speaking ability. The central question was, "Do you think this speaking test will change your study habits toward the improvement of your speaking ability?"

Results

Reliability. The interrater reliability was acceptable (over .74 among ten raters). In addition, a reasonably high correlation (the range was .74-.90 between individual native English speaking raters and Japanese raters) was obtained. This fact indicated that Japanese teachers by themselves can conduct the test and score the results in a classroom setting with little help or even without any help of NSs (within the reliability range of .74-.90). An internal consistency reliability estimate (over .84 for 10 raters) was obtained, and it indicated that the items were measuring the students' speaking ability fairly consistently.

Task correlations. The task correlation results demonstrate a strong relationship between Task I (Speech Making) and Task II (Visual-Material Description), and the tight relationship between Task III (Conversational Response Activities) and Task IV (Mini Contexts). The two pairs of strongly related tasks were apparently playing complementary roles with each other.

There was also a high task correlation (over .81) between Japanese and native English speaking raters. In other words, Japanese raters will be able to evaluate students' speaking ability in almost the same way as native English speaking raters. This task correlation also supported the construct validity of the test in that speaking ability consists of four partially divisible tasks, which are theoretically motivated.

Factor analysis. Through factor analysis, I was able to obtain two factors: Factor 1 (Linguistic Ability) and Factor 2 (Interactional-Sociolinguistic Ability). Apparently, Linguistic Ability is measured primarily by two tasks (Speech Making and Visual-Material Description) and Interactional-Sociolinguistic Ability is found in the other two tasks (Conversational Response Activities and Mini Contexts).

Discussion of the Validity Findings

Construct validity was examined through factor analysis and task correlations. Through factor analysis, two factors (Linguistic Ability and Interactional-Sociolinguistic Ability) were obtained. Task correlations further supported the construct validity of the test in that speaking ability consisted of partially divisible tasks (Speech Making Test, Visual-Material Description Test, Conversational Response Test, and Sociolinguistic Test). The two-factor structure, with the help of task correlations, partially supported the present author's proposed framework of speaking ability based on Bachman's Communicative Language Ability model.

Concurrent validity was investigated by comparing the results of the present test with students' grades in English Conversation classes and a teacher's estimate of students' speaking ability. The concurrent validity of this test was supported not by the class grades but by the teacher's estimate. This is probably because students' grades include non-language proficiency elements such as attendance, effort, and participation, among other things.

Washback validity was examined through a questionnaire administered to the students. Students' responses suggested that they will change their study habits (a) by focusing more on the productive aspects of their language skills, and (b) by paying more attention to the context, etc. It is hoped that this change will reach teachers, too, so that teachers will modify their teaching styles from grammar-oriented classes to communication-oriented classes.

Face validity was partially supported by the present author's informal talk with students. They were excited about taking this type of unfamiliar, but seemingly authentic, speaking test. Therefore, they were highly motivated to speak out in the testing situation.

Conclusion

In studying our language tests, we should strike a good balance of the various kinds of validity depending on the situation in each institution. However, by far the most important validity is construct validity. Classroom teachers must always consider whether they are really measuring what they intend to measure. There are at least three ways to look at the construct of speaking: (a) the nature of speaking, (b) the theoretical or linguistic underpinnings of speaking, and (c) classroom teachers' ideas based on their teaching experiences. As Weir (1993) says, tests should be theory driven. The theoretical part is indispensable; nevertheless, the viewpoints of experienced teachers should also play a significant role in the process of test construction. By alternating back and forth between the theoretical and practical aspects of a speaking test, classroom teachers can always focus on what they are actually measuring.

Note

[1] Eventually 10 raters' (four Japanese and six Native) results were used for statistical analyses.

References

APA (American Psychological Association). (1985). *Standards for educational and psychological testing.* Washington, D.C.: American Psychological Association.

Bachman, L. F. (1990). *Fundamental considerations in language Testing.* Oxford: Oxford University Press.

Davies, A. (1990). *Principles of language testing.* Oxford: Basil Blackwell.

Heaton, J. B. (1989). *Writing English language tests.* London: Longman.

Hughes, A. (1989). *Testing for language teachers.* Cambridge: Cambridge University.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.). *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education/Macmillan.

Nakamura, Y. (1993). *Measurement of Japanese college students' English speaking ability in a classroom setting.* Unpublished doctoral dissertation, International Christian University, Tokyo.

Spolsky, B. (1975). Language testing—The problem of validation. In L. Palmer & B. Spolsky (Eds.), *Papers on language testing 1967-1974* (pp. 147-153). Washington, DC: TESOL.

Thrasher, R. H. (1984). Educational validity. *Annual reports, International Christian University, 9,* 67-84.

Weir, C. J. (1990). *Communicative language testing.* London: Prentice-Hall.

Weir, C. J. (1993). *Understanding and developing language tests.* London: Prentice-Hall.

Section V

Innovative Testing

Chapter 15

Negotiating a Spoken-English Scheme with Japanese University Students

JEANETTE MCCLEAN
TOKYO DENKI UNIVERSITY

Since I am a teacher of English "conversation" in a private Japanese university, my students are typical: aged between 18 and 22, they follow a four-year degree course for which English is not a major subject. Most are false beginners (i.e., learners previously exposed to English instruction but having weak aural skills). Part of my role is to set examinations for these students and award a final year-end grade. Until the start of the project described in this chapter, the students' oral English ability, like many others at universities throughout Japan, had been measured through reading or writing tests. The likelihood was that this practice would continue, despite warnings about the harmful effect of such indirect tests, which encourage candidates to develop their ability to handle indirect rather than realistic tasks (Hughs, 1989; Brindley, 1989; Weir, 1990).

To continue to test oral English proficiency using indirect testing techniques in a country which now lays so much importance on spoken English proficiency is inappropriate. As Gruba (1993) states, "Test developers in Japan must seek to make examinations that are themselves responses to changes in the field" (p. 33). In the absence of any kind of framework for direct assessment of spoken English at this particular university, a research experiment was set up in 1992-93 to design and implement a direct test for assessing the oral ability of freshmen. This chapter is based on the action research conducted as a part of this project.

Problems in Assessing Spoken English at Japanese Universities

Before contemplating an innovation of this kind it is important to consider the context in which the new form of testing is to be implemented. By means of Heron's (1981) four-part assessment process, some of the problems that testers of spoken English in Japanese universities encounter are considered below.

What to Assess. In deciding what to assess, test developers face a fundamental dilemma about what actually constitutes oral competence in spoken English. Canale and Swain (1993) define communicative competence as an oral skill involving grammatical, sociolinguistic, discoursal, and strategic competence, while Bachman (1990) defines communicative language ability as "the ability to use language communicatively and involves both knowledge of or competence in the language, and the capacity for implementing, or using this competence" (p. 81). Weir (1990) stresses that, unless speakers are able to utilize the newly acquired language appropriately according to the demand of the situation, their knowledge of language effectively counts for nothing. Fairclough (1992), on the other hand, dismisses the whole idea of appropriateness as a sociolinguistic ploy and the "political objective of the dominant, 'hegemonic,' sections of a society in the domain of language as in other domains, but it has never been a sociolinguistic reality" (p. 49). Taking a more practical view of interaction, Weir and Bygate (1993) claim that competent speakers communicate by means of either informational routines or interactional routines—improvising, negotiating, dealing with communication problems as they arise, and managing the interaction, as well as controlling the topic, the development, and duration of the content.

While these notable theorists do hold in common the belief that communicative competence involves use of the language, appropriateness, and communicative strategies, the divergence of opinion about how this is manifested provides little reassurance for tester in Japanese universities, who are entirely responsible for deciding what they will teach to their students (Benson, 1991; Evanoff, 1993) and are also responsible for deciding how, and on what, students will be examined. It should come as little surprise, therefore, that what little oral assessment has been done in Japanese universities so far has been isolated, haphazard, lacking in form, and subjective. Testers are confused about what specifically to focus on when assessing test-takers' oral competence. The autonomous nature of teaching in Japanese universities and the lack of cohesion in testing spoken English is worrisome and exists despite efforts to work within a communicative paradigm and despite Monbusho encouragement to teach language communicatively.

Which Criteria to Use?

The second decision concerns which criteria to use for assessment. Testers are not only faced with the problem of what to assess, but must also face the immense cultural differences between English L1 speakers and Japanese speakers of English. Testers must ask themselves whether test-takers should be assessed on their "inner directed" behavior based on the social roles and models in Japanese society, or their "other directed" behavior based on Western, and particularly American, social norms. Alternatively, should the criteria adopted be more international and pertinent to all nonnative speakers of English, or specifically oriented to the profile of a Japanese university student learning English? Test-makers and test-markers (those who must mark and score the tests) face an agonizing dilemma over this issue alone.

For testers in Japanese universities, it is also important to bear in mind that the predominant method of grading learners is norm-referenced and has its roots firmly based in the traditional university system. The aim of the scoring procedure is to rank test-takers in relation to their fellow test-takers by means of the score or percentage attained. Unfortunately, this evaluation method bears little relation to the learner's competence in the target language. However, alternative criteria for language assessment based on performance objectives are being adopted more and more, utilizing techniques which seek to focus on the learners' mastery of the language. Criterion-referenced tests (CRT) evaluate an individual's performance of specific communicative skills (Brown, 1988), and provide information directly about what the learner can actually do with the target language. If testers in Japanese universities choose to adopt this method of assessing their students, not only must they be able to identify the skill inherent in spoken English, they must then create or select the appropriate criteria against which to assess their testees' performance. Such choices are made all the more difficult by the differences in cultural perspectives mentioned above.

How to Apply the Criteria

The next decision is concerned with how to apply the criteria selected. For those wishing to adopt a more performance-based approach to assessment, the difficulty of converting a CRT-oriented assessment scheme, which may involve several scores, to the required single grade is strained by the very structure of the grading system used in Japanese universities—a system which is not negotiable. In a fairly typical Japanese university, an "A" grade may be awarded for a score of 80% and above, a "B" for 60%-75%, a "C" for scores within a 5-point range around 60%, and a "D" (failure) for a score below 55%[1]. A fundamental flaw exists in a grading structure of this type because the number of scores possible within the top and bottom grades (i.e., "A" and "D") are heavily weighted, while the possible number of scores possible in the middle range (i.e., grades "B" and "C") is limited. A graph plotted to show the number of scores possible within the given grade ranges shoes a distribution in a U-shaped curve—the inverse of a normal distribution. A normal distribution—a bell-curve—assumes the majority of scores fall within the middle range of scores (i.e., grades "B" and "C") while high and low scores are rare. Perhaps this U-shaped structure might account for some of the unusually high number of "A" grade graduates produced by Japanese universities.

A paradox exists within the system: the importance of test results can be challenged by teachers. Teachers are able to dismiss student test results through experienced judgment, intuition, or for other reasons. Testers are rarely held accountable for their methods of grading. This, together with the absence of uniform test criteria, serves to show the power that teachers have to influence students' progress in the educational context (Heron, 1981).

In addition to the authority exercised by staff, Japanese university students are discouraged by their peers from displaying their oral ability in English. Inculcating a communicative teaching approach in a Japanese university is difficult because of the pressure to conform, rather than to display one's individual feelings, opinions, or personality in Japanese society (Nozaki, 1993). Japanese students are not trained in critical thinking and debate because social training has reinforced passive submission to authority (Jones, 1991a, 1991b; Nozaki, 1993; Davidson, 1994) where, to quote a well-known Japanese saying, "The nail that sticks up gets hammered down." The system offers little incentive for improving personal performance, a view apparently supported by the prevailing belief amongst many Japanese university students that the university is an "escalator system" where, "if both feet are placed in the bottom step, a student almost automatically progresses to graduation" (Nozaki, 1993, p. 28).

Doing the Assessment

For testers in Japanese universities the fourth part of the assessment process—actually doing the assessment—may well be like charting a course through a minefield. Time constraints for the completion of student assessment are one of the biggest obstacles to achieving reliable testing, and these are seldom negotiable. Add to this the limited number of staff and accommodation available for test administration, and circumstances dictate that the teachers/assessors must add a third role to their repertoires—that of test administrator.

The situation is compounded by one more factor, one that Jones (1992) cites as the loudest complaint of foreign staff in Japanese universities: large classes. In real terms, direct assessments of speaking ability, such as interviews, are time-consuming and difficult to administer if there are large numbers of candidates (Weir, 1990). In Japanese universities, it is commonplace for testers to be required to conduct English tests for 50 to 120 students within 90 minutes, and to have completed marking and grading from 200 to 600 students within one week prior to the examinations for "regular" subjects. Apart from the difficulty that such numbers impose, the likelihood of obtaining realistic discourse samples from the test-takers is hampered because the processing of interchanges under normal time constraints, as advocated by Weir (1990) and Morrow (1979), is simply not feasible under these conditions.

The staff shortage and large number of test-takers poses the additional problem of test security: how to prevent the students passing information on to each other. Fortunately, the very nature of spoken interaction is unpredictable, reciprocal, and adaptive (Bachman, 1990), which, for the student, means that it is difficult to replicate natural, interactive speech. For the assessor, these characteristics of spoken English mean that it is difficult to make tests for spoken English reliable. Weir (1990) stresses the importance of familiarity from the test-takers' point of view for a test to be acceptable; but Japanese university students are not used to being assessed orally and, despite careful briefing and preparation of candidates, it is difficult to make an oral performance test reliable in this context by limiting the element of unpredictability to the interchange alone. Furthermore, previous research has shown the difficulty of achieving marker consistency in assessing student oral performance in English (McClean, 1993; Nambiar & Goon, 1993).

Hence, in this context, testers are restricted in the form of assessment they can select to one that is predictable rather than desirable (Weir, 1990). This may, in part, explain the reticence thus far in Japanese universities for introducing direct tests to assess spoken performance in English, and the retention of entrance examinations with a strong bias towards grammar, reading, and writing skills (Benson, 1991) with no oral component (Jones, 1991a, 1991b).

Reasons for Initiation of the Testing Project

While acknowledging some of the more obvious problems affecting the introduction of direct testing of spoken English inherent in the present Japanese university system, I am ill-at-ease in the knowledge that unilateral authority of the teacher and unreliable evaluation methods leave students without a clear impression of their performance in spoken English. Learners are, instead, subject to the politics of knowledge present in an educational model which is rigid and highly authoritarian. I concur with Heron's (1981) observation that "...The time is ripe for an alternative, democratic model: that of equal human capacities which mutually support and enhance each other" (p. 61).

Kelly (1993) points out that the role of the university in Japanese society is the reverse of that seen in Western society. Whereas in the West the university phase is one of narrow specialization, in Japan university is seen as a broadening phase important for personal development, cultivating independence, and human relations. However, Japanese university students lack skills in critical thinking, they are transient, and they have only limited contact time with their teachers. Surely what we, as language teachers, should be doing is offering our students opportunities to experience different learning styles and providing valuable broadening experiences in which the learners can rehearse a degree of autonomy in their maturation process. Collaborative assessment offers one such avenue because it provides an intermediary stage between self-determination and unilateral assessment. As Kelly further notes,

> Traditional Japanese patterns of interaction and problem solving are powerful in their own right, but limited: they are still rooted in an agricultural value system with a predisposition toward constancy rather than change. (p. 185)

Because of the low level of English, weak fluency skills, and limited amount of spoken language produced by Japanese university students in any given interactive situation, evaluation of their oral proficiency is very difficult for an assessor. It seems appropriate, therefore, that what English language teachers should now be seeking is a model of education in which intellectual, emotional, and interpersonal competence go hand in hand.

The democratic model generates a set of guiding norms for the management of feeling (Heron, 1981). To encourage the opening of feelings in Japanese students is not only very important because the L1 culture actively discourages the display of emotion, but also because they are learning a language in which communication is based on the intention of the speaker. For Japanese learners, English is a counterculture and provides much of what Japanese culture withholds: the sense of one's uniqueness, direct expression of one's opinions, and a focus on content rather than form (Kelly, 1993). Furthermore, English offers an alternative mindset for dealing with modern problems by offering a greater variety of responses to the increasingly diverse situations modern Japanese are faced with. In many ways, Japanese university students are not yet ready for oral assessment of their spoken English skills. If university level is the phase when Japanese students are expected to develop self-confidence, individuality, and a clearer understanding of life, then a teacher-dominated, lecture-test approach is inappropriate and counterproductive at this stage.

The outcome of two initial trials to develop a suitable test and grading scheme at this university showed that achieving reliability in terms of test format and marker consistency were simply impossible to achieve in view of the problems described above. It therefore seemed more productive to switch the emphasis in testing at this university from the product, i.e., "what" language the students had learned or produced, to the process, i.e., "how" one goes about communicating in the L2. Heron (1981, p. 64) advocates that process assessment is more important than content assessment because "…procedural competence is more basic than product competence, since the former is a precondition of providing many good products."

We decided, therefore, to focus on the development of a new grading scheme through which students could be familiarized with the techniques and skills involved in the process of oral communication. To be of any real benefit to participants in the testing process, however, it was essential for all parties to negotiate and agree on assessment criteria, against which the performance of the freshmen would be assessed in their final examination. The process employed to arrive at a **Negotiated Grading Scheme** (NGS) is described below.

Developing a NGS: The Process

Step 1—Brainstorming. This involved searching for a common point of reference from which to start the collaborative process. As their teacher, I was concerned to find out what the freshmen actually understood about speaking ability and the skills involved in spoken English. When asked to describe a good speaker of English, it was distressing to discover that, after completing two-thirds of a course which emphasized the skills of spoken communication, the students were unable to even attempt a description of a good speaker of English.

This, then, was the starting point. I decided to present the freshmen with examples of good, fair, and weak English speakers from amongst the target test population to see if they could distinguish between speakers' proficiency. Two videos were presented of two pairs of students performing an information gap task (IGT) under examination conditions. After viewing the videos, the freshmen were asked to comment on whether they thought the speakers were good, fair, or weak English speakers. They responded that some candidates' performances were fair but others were "more fair"!

To give the freshmen the opportunity to focus more carefully on the first pair of speakers, half of the class were asked to observe Speaker 1 (S1) and the other half to observe Speaker 2 (S2). After viewing the speakers again, the students wrote three things their speaker was good at and three he was poor at. For instance, "S1 has clear pronunciation, he can use long sentences, and he appears confident and relaxed; however, he doesn't make eye contact or try to help his partner." Gradually, descriptions of the speakers' performances were developed by eliciting characteristics from the students and collating these on the board. The same process was repeated with the second pair of speakers. After viewing the second IGT, discussion followed about what the freshmen considered to be a good, fair, and weak speaker of English. For homework they were asked to write a short profile of these three levels of speaker.

Step 2—Correlating and Condensing. The next step was the development of a table of descriptors (Table 1). Descriptions were collected from the students and, together with input from the teachers, a comprehensive table describing the skills involved in spoken English was drawn up by the teachers.

To honor input from all participants (in the spirit of cooperation), it was important to retain the terminology used by the freshmen as much as possible. Students' responses were correlated and simplified into short descriptions. To simplify comparisons, wherever possible descriptors were matched across the three levels by adding intensifiers. Single descriptors considered important were also included in the table and placed under what the teachers considered the appropriate level.

It was interesting that amongst the criteria freshmen indicated as important in speaking English were qualities such as confidence, a positive attitude to learning, and characteristics which might be considered non-communicative. Brindley (1989) supports the practice of including non-communicative criteria in CRT-assessment if they are considered important. It is also interesting that Jones (1991a, 1991b) revealed from his survey of English language teachers in Japanese universities that these are the very same qualities teachers are most concerned to boost in their learners.

Table 1. *Descriptors for Speakers of English*

A Poor Speaker	A Fair Speaker	A Good Speaker
lacks confidence	tries to communicate	is confident
is nervous	is relaxed	
is shy		
is afraid to make mistakes	is a little afraid to make mistakes	is not afraid to make mistakes
makes no eye contact	often makes eye contact	makes good eye contact
does not use gestures	tries to use body language	uses body language
uses body to communicate, not words		
speaks in a quiet voice	sometimes can't be heard	speaks in a loud voice
pronunciation is not clear	sometimes makes pronunciation mistakes	has clear pronunciation
has a small vocabulary	uses easy words	has a large vocabulary
speech is not natural	sometimes speaks smoothly	speaks smoothly
has flat intonation	has some intonation	speaks expressively
thinks a long time between words, makes long silences	tries to speak fast but makes pauses often	can speak fast or slow
is hard to understand	tries to speak but is sometimes hard to understand	is easy to understand
	is not perfect but listener can understand	helps other speakers
can't describe what he wants to say	can describe what he wants to say but it is not easy	communicates well with the listener
can sound rude		sounds polite

Step 3—Jigsaw Exercise. Step 3 was an important consultative phase in the collaboration process, when the descriptors were taken back to the freshmen for approval. To encourage them to think critically, the table of descriptors was cut up into individual descriptors, placed in an envelope and given to small groups of students. First they were asked to arrange the descriptors under three headings, thereby developing profiles of good, fair, and weak English speakers. They were invited to discard any descriptors they felt were irrelevant or inappropriate, or add any they felt should be included and had been omitted.

In the second task, the freshmen

Figure 1. *Spoken English Ability Performance Profile*

Name: ______________________ Student No. ______________

English Oral Examination, January 1995

Performance Profile

Points	Attitude & Confidence	Body Language	Expressiveness (pronunciation, intonation & volume)	Understand ability (for the listener, is the message delivered clearly?)	Communicative ability (can the speaker say what he/she wants to say?)
Good 5					
4					
Fair 3					
2					
Weak 1					
				Total: ______ 25	

were asked to group the descriptors covering similar skills or characteristics into categories across the three ability levels (e.g., descriptors about nonverbal communication, expressiveness, etc.). Once this was completed, the freshmen were asked to rank the categories in order of importance. In the final task, they were asked to give a title to each cluster of descriptors.

Clearly this task was the most difficult for the students. Some interesting titles emerged, such as "Attitude/Heart/Emotion"; "Body Language/Gestures/Action"; "Understanding/Listener's think/Easy to understand"; "Knowledge/Vocabulary/Words/Communicativeness"; "Confidence/Relaxation/Feeling/Spirit," revealing the possibility of significant cultural differences in perception between participants about various elements of speech. Interestingly, very similar terminology was elicited on a separate occasion from Japanese English language teachers when performing the same exercise, which serves to show the importance of this step in arriving at mutually acceptable criteria and terminology.

Step 4—Drawing up the Grading Scheme. Again the input of the freshmen was correlated by teachers, and two rating scales were developed: a Performance Profile and a Description of Performance.

1. *The Performance Profile.* In keeping with the CRT-orientation of communicative testing, the Performance Profile (PP) (see Figure 1) sought to present an individual profile informing freshmen about their individual mastery of certain skills of spoken English. The five most important categories identified by the freshmen were selected for inclusion into the PP, the underlying components of which are the criteria identified by participants in Steps 1, 2, and 3 above. The categories are represented on the horizontal axis of the assessment grid.

In response to the students' initial dissatisfaction with the distinctions "good", "fair", and "weak" speaking ability, two additional levels were included in the vertical axis to fudge the distinctions, but for the sake of consistency these original terms were retained. Five bands of ability, ranging from

Figure 2. *Description of Performance*

Very Good Speaker
(21-25 points)

Is very confident and relaxed; maintains eye contact and uses body language well; speech is clearly heard and pronunciation is clear and accurate; speaks expressively with varied intonation; seldom makes pauses and speaks fast or slow with ease; has a large vocabulary and can say what he/she wants to say with ease; helps other speakers; makes relevant replies.

Good Speaker
(16-20 points)

Is confident and not afraid to make mistakes; often uses body language and makes eye contact; usually speaks in a clear voice but occasionally mispronounces; is usually expressive and often sounds polite; usually speaks smoothly but occasionally rephrases; vocabulary is adequate and can describe what he/she wants to say; listener usually understands; replies are usually relevant.

Fair Speaker
(11-15 points)

Tries to communicate but is a little afraid to make mistakes; uses body language and frequently makes eye contact; sometime can't be heard; frequently mispronounces; intonation is fairly expressive but occasionally sounds rude; makes pauses often, sometimes speaks smoothly; uses easy words, can describe what he/she wants to say but it is not easy; is frequently hard to understand; occasionally replies are off the point.

Developing Speaker
(6-10 points)

Often lacks confidence, is afraid to make mistakes; uses many Japanese gestures, occasionally makes eye contact; pronunciation is very Japanese and often unclear; intonation is flat and often sounds rude; speech is not smooth and has long pauses; has a small vocabulary and often has difficulty describing what he/she wants to say; is often difficult to understand; replies are frequently off the point.

Weak Speaker
(1-5 points)

Lacks confidence, is nervous, shy and often does not communicate for fear of making mistakes; uses body language to communicate not words, does not make eye contact; speaks quietly, pronunciation is unclear, intonation is flat; speech is slow and makes long silences; has a very small vocabulary and cannot describe what he/she wants to say; is very difficult for listener to understand; replies are often off the subject.

weak (level 1) to very good (level 5) and represented on the vertical axis, were included to allow the marker a wider scope and more flexibility in marking. Test-takers would be awarded a score of between 1 and 5 according to their performance in each category and given a total score out of 25 for their performance in all five categories.

2. *The Description of Performance.* In contrast to the PP, the Description of Performance (DOP) (see Figure 2) is a holistic scale and attempts to describe the learner's overall performance as an English speaker. In this respect, the DOP resembles a norm-referenced type of grading scheme. On this scale, each level is equivalent to a range of ability (i.e., weak, developing, fair, good, very good). For each level, a description of a model speaker was drawn up from the table of descriptors elicited from the freshmen in Steps 1, 2, and 3.

In formulating these two rating scales, it was anticipated that the test-taker would first be assessed on the PP and then separately on the DOP. The assessor would then compare the total score on the PP and the score band on the DOP to check that the scores were equivalent. Where the score given by the assessor on the PP did not fall into the equivalent band on the DOP, the assessor would need to reconsider the PP score, or remark the test-taker. In this way, the DOP acted as a check to ensure consistent scoring.

Step 5—Trialing the Rating Scales. The negotiated Grading Scheme (NGS) was trialed by both the freshmen and two testers. In the first trial, freshmen were shown pairs of university students performing an information gap task on video and were asked to rate one speaker's performance on the PP in one category only ("Expression"). Afterwards, scores were compared and difficulties experienced in marking were discussed. The process was repeated, but this time, freshmen were asked to rate the speaker's "Attitude/Confidence." Scores were again compared and discrepancies discussed. After the third showing of the video, the freshmen were required to mark the speaker's performance in the remaining categories on the PP.

After discussion about the scores, the students were asked to give an overall rating of each speaker on the DOP before totaling the score on the PP. This final step operated as a necessary brake on the marking process, and an important precaution. It made the evaluators stop and think critically about the speaker's overall performance and discouraged them from making an immediate comparison between the PP and the DOP scores. From the equivalence or discrepancy between scores on the two rating scales, the evaluators could see how consistent they were in assessing their peers' performance.

The NGS was also trialed by two teachers who utilized the two scales for assessing freshmen in their final oral examination. The test involved groups of four students, who were given 15 minutes to discuss a picture or photo of their choice. Each test-taker was required to ask and answer a minimum of one question in the discussion. These test-takers were assessed simultaneously by the assessor, but the interactions were also recorded on video.

Bearing in mind that the test format used was not necessarily the most desirable, but was most practicable under the circumstances, the grading scheme proved easy to use for the assessor as it facilitated quick and simultaneous marking of 4 candidates within 15 minutes. Only occasional remarks were necessary, and these were made possible by the video recordings. In this way, the testing schedule was not disrupted. Interestingly, the assessors endorsed the freshmen's finding that the PP was easier to use for marking than the DOP, but felt that the DOP acted as an important check against which to confirm the PP score. Results showed that random rescoring of these test-takers by the assessor indicated a very high degree of marker consistency (99 percent using the Pearson Product-Moment Correlation method).

Step 6—Feedback to Students on Their Results. After their examination, during the final class period of the year, freshmen were invited to collect their PP and DOP from the assessor and informally discuss their results. Most students were concerned to know how they had performed and, although some were disappointed with their results, they gained a more realistic idea of their performance under exam conditions. As their teacher, I was able to indicate discrepancies between classroom and examination performances and give freshmen a clear idea about which skills to focus on in the future to improve their spoken English.

Benefits To Participants

All participants agree that several important benefits for the students emerged from this project. On a questionnaire completed just prior to their examination (n=104), 64% of the freshmen stated that they enjoyed being involved in the negotiation process, 73% responded that as a result of this process they understood what skills a good speaker of English displays, and 70% felt that the descriptions in the grading scheme were clear to follow. The collaborative process was helpful, as 56% of the freshmen reported that they did not know their own strengths and weaknesses as English speakers before their course began. Many also stated that they were now less afraid of making mistakes. Before the examination, 61% of the test-takers believed that knowing how they were going to be assessed would improve their performance in the examination—a view which, I believe, indicates that the majority of freshmen had taken ownership of this grading scheme before the test took place.

From this feedback, it can be concluded that the collaborative approach increased the learners' confidence and reduced anxiety as, by the time they performed the examination, the freshmen knew what they were going to be assessed on. As an exercise in critical thinking, the development of this grading scheme encouraged individual initiative and gave freshmen a sense of control over a learning environment which for so long has controlled them. That 72% of respondents agreed that developing the grading scheme increased their desire to improve their spoken English shows the significance of consultation in fostering a high level of motivation amongst university students.

Although the collaborative process still allows the teacher to retain the power of veto, this experiment suggests that these particular learners are capable of setting their own standards and will become more responsible for their own progress in the future. In bringing about a shift in responsibility from the teacher to the learner, Japanese university students can be prepared for a change of teacher from year to year and, more importantly, for lifelong learning. By adapting this technique for peer evaluation in the classroom, rather than being a drawback for the learners, utilizing an NGS could actually be less threatening than the unilateral examiner/test-taker method and prove to be a more meaningful method of assessment for them. Such preparation is increasingly important for young graduates in Japan today, in view of the insecure and competitive job market they face on exiting the university. Japanese university students do, however, need to be prepared and trained in the collaborative process.

For the assessor, there were several benefits in utilizing this NGS. Firstly, a very significant degree of intra-marker reliability was achieved on the total scores, which shows a high degree of certainty and consistency on the part of the marker. This suggests that, not only is the process of agreeing on criteria beneficial for students, but it also helps clarify the criteria in the assessor's mind prior to marking. The inclusion of a second rating scale may also have raised intra-marker reliability because the DOP provided an immediate reference check, taking the pressure off the assessor to provide a "cut-off" score or be exact to one point on the PP. By instilling this element of flexibility, fatigue was reduced for the assessor, which is an important consideration for testers working under the strict time constraints imposed by Japanese universities. For instance, when the assessor remained undecided about a test-takers' performance, recording test-takers' interactions on videotape provided the means for an additional check without delaying the test schedule or recalling the test-taker.

The overall results of this experiment were also encouraging. The distribution of test-takers' scores obtained using the PP revealed a normal distribution (a bell shaped curve) with scores falling over a fairly wide range (range = 19 out of 25). This range, together with the standard deviation of 4.2, show a wide dispersion of scores, suggesting that the assessor has flexibility to award scores over a wider range using an NGS than a commercial grading scheme. In the initial trial to develop a suitable oral test, the TEEP2 grading scheme was utilized and, although scores were distributed over a wide range (range = 17 out of 20), the standard deviation was narrower (3.91) and a negatively skewed distribution of scores was obtained. Using the NGS, a median and mean of 13.5 was achieved, which was higher than that attained with the TEEP (median = 9; mean = 9.32). This suggests that an NGS is more suitable for use with this group of test-takers because on this scheme they can achieve a higher overall score, whereas the TEEP scheme is oriented toward all non-native speakers of English.

Relatively low correlations were obtained, however, in individual categories using the NGS (Understandability = 0.638; Intelligibility = 0.218), suggesting that some skills are more difficult to mark and/or that the tester was accommodating to Japanese learners and, in particular, to her students. This simply illustrates the necessity of a collaborative process to train assessors to a common standard, in order to attain a high degree of inter-marker reliability and the use of at least two markers to counter subjectivity in marking.

For the teacher, the main benefit of this CRT type of assessment scheme is that it provides information about learners' performances. Strengths and weaknesses can be isolated across the whole test population, allowing adjustments to be made in both curriculum and course content to accommodate the learners' needs. It also permits specific information to be gained about an individual's performance—a distinct advantage in a learning context typified by large classes and limited contact time. An NGS provides a tangible base from which the teacher can advise and cooperate with individual students, and together they can agree on attainable target levels of performance to aim for by the end of the following year. And, while there is the drawback of testing in Japanese universities that the assessor and teacher are usually one and the same person, this situation does allow discrepancies between learners' examination and classroom performances to be recognized.

Through this cooperative process, teachers are also prompted to think critically and become better able to judge the appropriateness of different forms of evaluation. This is an important point to consider in the existing light of English language departments in Japanese universities, which employ large numbers of part-time teachers and staff on short-term contacts who have varying proficiency and experience in the teaching and assessing of EFL/ESL.

Conclusion

The collaborative process used in this experiment increased trust between the students, teachers, and assessors because all parties' contributions were respected. Consultation built consensus, a vital pre-requisite for the acceptance of innovation in Japanese society, and the test and grading scheme achieved face validity in this context because they appeared to measure what they were supposed to (Hughes, 1989), i.e., competence in spoken English. There is no question that this grading scheme has much room for improvement, and further effort is required to arrive at a test with high degrees of content validity and test format reliability. But, given the lack of awareness about communicative competence, a pervading fear of change, and regrettably little desire for inculcating communicative forms of assessment in Japanese universities, the problem, as Gruba (1993) puts it, is that "An understanding of the inability of tests to serve as perfectly neutral, perfectly effective instruments to inform decisions of resource management is crucial here" (p. 4).

Developing an NGS to fit the particular test population is an important step towards achieving political acceptability (Henning, 1987) for a test of oral English in Japanese universities. It opens up the mysticism of assessment, making test criteria transparent, and in so doing encourages ethical marking by assessors (Grube, 1993) and conformity to stated criteria, holds teachers accountable for the quality of their courses, and gives students the benefit of more effective teaching and evaluation. A negotiated grading scheme is, therefore, a worthwhile tool for teaching, and as a pre-requisite to the development and introduction of alternate forms of evaluation in Japanese universities out of which, it is hoped, more reliable and valid tests of spoken English may evolve.

Notes

[1] The existing grading structure used at Tokyo Denki University.

[2] The Test in English for Educational Purposes from the Associated Board, taken from Weir, C. J. (1990) Communicative language testing, p. 147-148. For the oral section of the TEEP test criteria focus on six categories of communicative competence (Appropriateness, Adequacy of Vocabulary, Grammatical Accuracy, Intelligibility, Fluency, and Relevance and adequacy of content). In each category four levels or bands of ability are described. Performance is rated from 0-3 according to the test-taker's fulfillment of the description in each band. A total score out of 20 is awarded.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Benson, M. (1991). Attitudes and motivation towards English: A survey of Japanese freshmen. RELC Journal, 22(1), 34-48.

Brindley, G. (1989). *Assessing achievement in the learner-centred curriculum*. Sydney: National Centre for English Teaching.

Brown, J. D. (1988). *Understanding research in second language learning*. New York: Cambridge University Press.

Canale, M., & Swain, M. (1980). Theoretical basis of communicative approaches for second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Davidson, B. W. (1994). Critical thinking: A perspective and prescriptions for language teachers. *The Language Teacher, 18*(4), 20-25.

Evanoff, R. (1993). Making a career of university teaching in Japan. In P. Wadden (Ed.), *A handbook for teaching English at Japanese colleges and universities* (pp. 15-22). New York: Oxford University Press.

Fairclough, N. (1992). The appropriacy of "Appropriateness." In N. Fairclough (Ed.), *Critical language awareness* (pp. 35-56). London: Longman.

Gruba, P. (1993). *Student assessment final report*. Chiba, Japan: Kanda Institute of Foreign Languages.

Henning, G. (1987). *A guide to language testing*. Cambridge, MA: Newbury House.

Heron, J. (1981). Assessment Revisited. In D. Boud (Ed.), *Developing student autonomy in learning* (pp. 55-68). London: Kogan Page.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Jones, J. B. (1991a). The FCET survey. *Bulletin of Joetsu University of Education, 10*(2), 223-224.

Jones, J. B. (1991b). The FCET survey. *Bulletin of Joetsu University of Education, 11*(1), 159-174.

Jones, J. B. (1992). The FCET survey. *Bulletin of Joetsu University of Education, 11*(2), 263-273.

Kelly, C. (1993). The hidden role of the university. In P. Wadden (Ed.), *A handbook for teaching English at Japanese colleges and universities* (pp. 172-191). New York: Oxford University Press.

McClean, J. M. (1993). An experiment in test design for assessing the English oral ability of TDU freshmen 1992-3. *Research reports of the Faculty of Engineering General Education, Tokyo Denki University, 12*, 43-50.

Morrow, K. (1979). Communicative language testing: Revolution or evolution. In C. J. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching*. Oxford: Oxford University Press, 143-158.

Nambiar, M. K., & Goon, C. (1993). Assessment of oral skills: A comparison of scores obtained through audio recordings to those obtained through face-to-face evaluation. *RELC Journal, 24*(1). 15-31.

Nozaki, K. N. (1993). The Japanese student and the foreign teacher. In P. Wadden (Ed.), *A handbook for teaching English at Japanese colleges and universities* (pp. 27-34). New York: Oxford University Press.

Weir, C. J. (1990). *Communicative language testing.* London: Prentice Hall International.

Weir, C. J. & Bygate, M. (1993). Meeting the criteria of communicativeness in a spoken language test. Paper presented at RELC Conference on Testing and Evaluation, Singapore.

Chapter 16

Assessing the Unsaid: The Development of Tests of Nonverbal Ability

NICHOLAS O. JUNGHEIM
RYUTSU KEIZAI UNIVERSITY

As the definition of what constitutes communicative competence is expanded and refined, our view of the learner's needs in the foreign language classroom have also changed considerably. While classroom teachers have attempted to deal with these developments by introducing more communicative teaching methods, appropriate instruments are not always available to test the effectiveness of these methods. How do we know if the learner has really acquired the new communicative tools that the teacher has exposed them to?

The purpose of this chapter is to show how standard test construction methods can be applied to the development of tests which address abilities not usually considered in traditional language testing. This chapter will describe a theoretical framework for testing nonverbal behavior and the construction and administration of two tests of nonverbal ability: the Gesture Test (Gestest) for assessing the comprehension of English gestures and the Nonverbal Ability (NOVA) Scales for assessing nonverbal behavior in conversations.

Background

With the expansion of the communicative competence framework (Canale & Swain, 1980; Canale, 1983), efforts have been made to improve oral proficiency assessment using the communicative paradigm. Along these lines, Bachman and Palmer (1983) developed the *Oral Interview Test of Communicative Proficiency*, which includes scales for measuring grammatical competence, pragmatic competence, and sociolinguistic competence as part of what Bachman (1988, 1990) refers to as Communicative Language Ability (CLA), his testing model of communicative competence.

While Bachman and Palmer's test was a major step away from previous tests which have been criticized as discrete-point tests without a basis in theory or research (Bachman, 1988; Bachman & Savignon, 1986), it relied solely on the vocal channel for its evaluation of oral proficiency, in spite clear indications that nonverbal behavior is also a part of the sociolinguistic and strategic competence components of communicative competence (Canale, 1983; Canale & Swain, 1980). While there have been ongoing calls to pay more attention to nonverbal behavior as an integral part of communication (Al–shabbi, 1993; Baird, 1983; Hurley, 1992; Kellerman, 1992; Pennycook, 1985; von Raffler-Engel, 1980), little has been done to address this issue for language learners. On the other hand, there has been an least one attempt to deal with the basic nonverbal communication skills necessary to do well at American colleges with the Communication Competency Assessment Instrument (CCAI) (reported in Rubin, 1982) affirming the importance of having nonverbal communication skills.

A further rationale for creating tests to evaluate language learners' nonverbal behavior is provided by research that has examined how ratings of oral proficiency are affected by what the raters see. In Neu's (1990) comparison of two learners' oral proficiency ratings and their use of nonverbalbehaviors, she found that raters were "fudging," or changing their ratings, according to their perceptions of each learner's nonverbal behavior. This resulted in a mismatch between oral proficiency ratings and actual linguistic data. The Japanese subject's scores were negatively affected by the raters' perception of his nonverbal behavior while the Saudi Arabian's were positively affected.

In their comparison of the face–to–face ratings of learners' oral proficiency with ratings of audio recordings of the same conversations two months later, Nambiar and Goon (1993) found that subjects received lower ratings on the subsequent audio recordings. Raters were "irritated" by long pauses and repeated grammatical and phonological errors which had been less noticeable in face–to–face ratings "because of the need to simultaneously focus on visual paralinguistic and extra–linguistic cues" (p. 24). Face–to–face assessment resulted in the raters giving the subjects more favorable ratings.

In addition to problems related to oral proficiency testing illustrated by the above examples, the simple gesture is a nonverbal cue which can cause confusion when different cultures meet (Jungheim & Ushimaru, 1990; Kimura, 1979). A simple wave of the hand to "come here" in Japan can also look like a Japanese person is waving hello or goodbye; the American just smiles and waves back. An important tool for strategic competence, or getting the message across by all means available, is missing in this intercultural exchange if the American is not aware of the meaning of a simple Japanese gesture.

The acquisition of gestures, however, varies even in L1. Kumin and Lazar (1974) found significant differences between three– and four–year–olds in their ability understand and use gestures included in a 30–item list of gestures called emblems, gestures that have a direct translation and meaning known to all the members of a given culture and are unambiguous even if taken out of context (Ekman, 1976) such as the "come here" gesture. They also found that both age groups were significantly better at understanding these gestures than using them.

Similar evidence exists for L2 gestures. Mohan and Helmer (1988) compared the understanding of English gestures by four– and five–year–old native English speakers and nonnative speakers and found significant effects for age and culture on the ability to understand an inventory of 36 emblems and illustrators, gestures used to underline, emphasize, or illustrate spoken language.

There is also evidence that the effectiveness of instruction in nonverbal communication in the foreign language classroom also depends on the learning style of the students. In his study of the classroom acquisition of emblematic gestures, Jungheim (1991) found that Japanese students who received instruction using a traditional presentation, practice, and production approach did significantly better on a posttest than students who were taught the same gestures using a more "communicative" approach.

The above discussion has shown some of the issues involving nonverbal communication as it pertains to language learners, but until now, there has been no theoretical framework to help language teachers deal with it.

Testing Fragment

A basic framework for the development of tests of nonverbal communication was first presented in Jungheim (1994d). This framework expands Bachman's (1990, 1991) Communicative Language Ability (CLA) model to include a three–part nonverbal ability component. Nonverbal ability can be defined here as knowing how to use and interpret a variety of nonverbal behaviors or cues appropriately for the target language and culture. The three parts of this framework are nonverbal textual ability, nonverbal sociolinguistic ability, and nonverbal strategic ability.

Textual Ability

This ability is referred to as textual because in the CLA model (Bachman 1990, 1991) textual competence includes written and spoken language. Nonverbal textual ability, then, will also include "ways in which interlocutors organize and perform the turns in conversational discourse . . ." (Bachman, 1990, p. 88). The primary nonverbal behaviors under textual ability are gestures, head nods, and gaze direction when used to facilitate the interaction process as in backchanneling or turn–taking signals. Facial expressions such as smiles and frowns can also be included under textual ability.

Textual ability can be interpreted in terms of both the frequency and appropriateness of a person's use of the nonverbal behaviors concerned. When Japanese use head nods as a backchannel signal, for example, they have been found to nod both more frequently and at different times than Americans when speaking English (Maynard, 1987, 1989, 1990). This is related to what is called *aizuchi* in Japanese. Nonverbal and otherbackchannel signals such as "uhuh" or "yeah" are more likely to occur at the point of grammatical completion for Americans, whereas, for Japanese who frequently use *aizuchi*, they are also used by the listener at the speaker's pauses as a kind of encouragement to continue speaking. When and how often someone nods, therefore, can be used to assess their nonverbal textual ability.

Gaze direction can also be viewed in terms of frequency and appropriateness. Japanese, for example, can be expected to have different gaze behavior from Americans (Barnlund, 1989; Hattori, 1986). Asians in general have been found to focus less on the face or head of a speaker (Watson, cited in Argyle & Cook, 1976, p. 27).

Gaze direction, however, is consistent among members of a particular culture (Greenbaum, 1985; Jungheim, 1994b, La France & Mayo, 1978). Therefore, differences among cultures may often become noticeable when someone is speaking a second or foreign language. Listeners also tend to look at their partners much of the time they are listening, but they have been consistently observed to look away prior to beginning a speaking turn (Duncan & Fiske, 1985). In the case of a language learner, this type of looking away to think prior to speaking also increases when they have less control of the foreign language. According to Bialystok (1990), effective control gives the impression of fluency or automaticity. Inappropriate use of head nods and gaze direction, thus, will affect the listeners impression of the learner's fluency. Frequency and appropriateness of gaze direction changes can therefore also be used to assess nonverbal textual ability.

Sociolinguistic Ability

Sociolinguistic ability includes the ability to recognize the appropriate use of nonverbal behaviors such as the less frequent use of head nodding by native speakers of English as well as the ability to use and interpret gestures that vary from culture to culture. It is related to CLA sociolinguistic competence as a sensitivity to differences in dialect or variety, to differences in register and to naturalness, and the ability to interpret cultural references and figures of speech. (Bachman, 1990, p. 95).

Previous tests of nonnative speakers' gestural understanding such as Jungheim (1991) and Mohan & Helmer (1988) are examples of assessing nonverbal sociolinguistic ability. These tests primarily covered gestures called emblems, or gestures which have a direct verbal translation that is understood by all members of the same group or culture. As many as 67 emblems such as the "come here" gesture have been identified as used by North Americans (Johnson, Ekman, & Friesen, 1975). The Profile of Nonverbal Sensitivity (PONS) (Rosenthal, Hall, Archer, DiMatteo, & Rogers, 1979), while not designed for language learners, is a good example of how a test of nonverbal sociolinguistic ability could be designed. In this test, subjects are asked to judge nonverbal behavior in video performances that have the sound distorted to make the speech incomprehensible while leaving the intonation patterns intact.

It goes without saying that language learners need nonverbal sociolinguistic ability not only to improve communication but also to avoid what could be agonizing nonverbal misunderstandings. A simple "thumbs up" gesture that has a good meaning in North America is the equivalent of giving someone the "finger" (an obscene gesture in the United States) in some cultures. Nonverbal sociolin-guistic ability not only enhances the language learner's communication but also helps to avoid embarrassing misunderstandings.

Strategic Ability

Nonverbal strategic ability is important for language learners because it covers the compensatory role of nonverbal behavior as well as its role in supporting and enhancing speech. This ability includes the learner's use of nonverbal behaviors such as gestures or mime to compensate for insufficient linguistic knowledge, when necessary, as well as the appropriateness of the learner's use of gestures to support or enhance speech.

Lack of the use of gestures does not, however, indicate poor strategic ability unless there is an unresolved linguistic deficiency or, on the contrary, the learner is perfectly understandable without gestures. On the other hand, nonverbal behaviors such as "nose pointing" and too frequent self-pointing for "me" (sometimes found among Japanese learners) would be rated lower for appropriateness for enhancing and supporting speech with gestures signifying spacial relationships or shapes.

Table 1 summarizes some of the productive uses of nonverbal behavior according to the three nonverbal abilities in the nonverbal ability framework. This list consists primarily of productive uses of nonverbal behavior and is not meant to be an exhaustive review of nonverbal behaviors. Rather, this is a listof some of the major nonverbal behaviors that could be of interest to language teachers as easily observable and ratable behaviors.

Developizng Tests of Nonverbal Ability

The following sections will outline the process of developing two tests of nonverbal ability and the results of their administration to groups of Japanese learners of English. The first test is the Gesture Test (Gestest) developed to measure language learners' nonverbal sociolinguistic ability to interpret English gestures. The second test is the Nonverbal Ability (NOVA) Scales developed to measure language learners' nonverbal textual ability and nonverbal strategic ability in conversations.

The Gestest

The Gestest described here is a completely redeveloped version of the test used in Jungheim (1991) as a research instrument. It consists of a female North American native speaker of English performing 30 gestures on video tape and a four–option multiple–choice answer sheet written in Japanese. The Gesttest was developed as a norm–referenced test using item facility (IF) and item discrimination (ID) as guidelines for good items. The development process basically followed Brown's (1995) three steps of (a) piloting a large number of items, (b) analyzing the items, and (c) selecting the best items to make up a smaller and more effective version of the test.

The following research questions frame this approach to the construction of a test of nonverbal sociolinguistic competence:

1. What gestures are important for language learners?
2. What is the difference in the understanding of these gestures between native and nonnative speakers of English?
3. What are the characteristics of a test of these gestures when applied to language learners?
4. How reliable and valid is this test as a measure of nonverbal sociolinguistic ability?

Table 1 *Use of nonverbal behaviors in a nonverbal ability framework*

Textual Use

Gestures

Hands are used by the speaker to emphasize speech.

Vertical head movement (nod) is used as a backchannel signal by the listener to indicate attention, understanding, or agreement.

Vertical head movement (nod) is used by the speaker as a within-turn or turn-end signal.

Horizontal head movement (shake) is used by the listener to indicate disagreement or with laughter.

Gaze

Listener-directed gaze is used at the end of an utterance to elicit a backchannel response.

Terminal gaze (prolonged gaze starting just before the end of an utterance) is used to signal the end of the utterance.

Speaker-directed gaze is used to signal attentiveness.

Facial Expressions

Smiles are used to indicate attention or agreement.

Frowns are used to indicate disagreement or lack of understanding.

Sociolinguistic Use

Gestures

Gestures unaccompanied by speech are used to convey specific meanings in a given culture.

Strategic use

Gestures

Mime (hand gestures) is used to compensate for a linguistic deficiency such as the lack of a necessary lexical item.

Hand gestures are used to support spoken language to communicate spatial relationships and physical shapes which are not always easily understood using spoken language alone.

Method

Subjects. The subjects in the Gestest development process were 14 North American native speakers (NS) of English (two females, 12 males) and 22 Japanese nonnative speakers (NNS), who were graduate students in TESOL at an American university in Japan. These subjects were used to pilot the gesture video. Subjects used to pilot the actual 30–item Gestest were 56 Japanese university students (12 females, 44 males) in three intact freshman English conversation classes at a Japanese university.

Materials. Materials consisted of a list of 54 gestures compiled from previous research (e.g., Jungheim, 1991; Kumin & Lazar, 1974; Mohan & Helmer, 1988) used to collect baseline data, a revised list of 38 gestures with an accompanying video, and a 30–item multiple–choice Gestest with a gesture video.

Procedures. The first step in the construc-

Table 2. *Gestest descriptive statistics*

Statistic	Pilot Video (NS)	Pilot Video (NNS)	Pilot Gestest	Revised Gestest
N	14	22	56	56
k	38	38	30	23
Mean	32.60	24.50	19.04	17.09
S	2.49	4.06	3.49	3.51
Median	32.00	24.50	20.00	18.00
Low	24.50	12.50	11.00	9.00
High	36.00	30.00	25.00	22.00
r	.87	.90	--	--
α	--	--	.63	.75
SEM	.90	1.28	2.12	1.76

tion of the Gestest was to determine which gestures to use for the collection of baseline data for the initial piloting. A list of 54 gestures was compiled from previous studies of gesture comprehension (e.g., Kumin & Lazar, 1974; Jungheim, 1991; Mohan & Helmer, 1988). The lists were given to three North American native speakers (NS) of English in Japan who were asked to (a) cross out those gestures that they thought would not be useful for language learners or they did not understand, (b) write simple descriptions of how the remaining gestures would be performed(examples of useful expressions for describing gestures were included), and (c) perform the gestures for a Japanese non-native speaker (NNS) of English who was instructed to write verbal equivalents for them in Japanese. Space was also included for adding any gestures that the three NSs thought were important.

As a result of the above elicitation task, 38 gestures that two out of three of the American NSs thought were important for language learners were chosen for the creation of a pilot video to collect basic item data. The gesture video was made by using a camcorder in a brightly lit room with a plain curtain for a backdrop. A female English teacher from the United States performed each of the gestures two times in succession while seated, according to the descriptions written by the above-mentioned three NSs. The video was then copied by connecting the camcorder to an ordinary video cassette recorder and adding numbers before each video performance. This video was shown to two groups of NS and NNS English teachers who were asked to write the meaning of each gesture. The NSs (14 North Americans) were asked to write the answers in English and NNSs (22 Japanese) in Japanese.

A Japanese and a NS English teacher then rated the answers using the gesture cues from the original 54-gesture list as a guide. These results were used to decide which gestures to include in the pilot version of the Gestest.

In order to determine how the understanding of each gesture differed between NSs and NNS, the item facility (IF) for each of the gestures was first calculated separately for NSs and NNSs. IF is simply the proportion of correct responses to a question on a test (number of correct answers divided by the total numberof persons taking the test). In this case, if 77 percent of the NSs correctly identified the gesture for "Okay," the IF would be .77, a relatively easy item.

Next, item discrimination (ID) was calculated for each item. Ordinarily, ID is used to see how well an item discriminates between a group of high scorers and a group of low scorers (see Brown, 1995 for a complete description of this method, or Chapter 5 for a brief description). In the present study ID was used to find out which gestures distinguished between NSs and

NNSs. The IFs of NNSs were subtracted from those of NSs for each item to see how well each discriminated between NSs and NNSs.

Using the IF and ID statistics for each of the 38 gestures, a list of 30 gestures was compiled for the construction of a pilot test. Simply speaking a good item should be very easy for NSs and have some degree of difficulty for NNSs. Gestures which all NNSs and NSs identified correctly, therefore, were also dropped. Gestures that were difficult for NSs were dropped, as well as gestures that were equally difficult for both groups. A number of these items were included, however, in spite of high IF and low ID statistics. The primary criterion for these items was that all of the NSs got the item correct or that the IF was below 80 percent for the NNSs. It was assumed that the high IFs for these items were a result of the English teachers' greater experience and longer years of speaking English. Such items/gestures were included to make sure that the pilot test had a larger number of items.

A new video was created using the original gesture video tape. Gestures were copied in random order preceded by a number. A four-option 30-item multiple-choice answer sheet was created with all options written in Japanese so that differences in English proficiency among the learners would not directly affect the results for gesture comprehension. The correct answers as well as the distractors were written using the answers written by the three Japanese persons who viewed performances from the initial 54-gesture list and answers for the pilot gesture video. These were arranged in random order for each item and reviewed by a native speaker of Japanese to assure the accuracy of the Japanese.

The resulting pilot Gestest was then ad-

Table 3. *Gesture Video and Pilot Test Item Facility*

Item	English Cue	Video Item	Video Pilot				Test Pilot			
			IF	NS	NNS	ID	IF	High	Low	ID
1.	Okay	(1)	0.87	1.00	0.77	0.23	0.68	0.88	0.31	0.56
2.	Me	(7)	0.87	1.00	0.77	0.23	0.77	0.94	0.56	0.38
3.	Yes	(19)	0.95	1.00	0.91	0.09	0.98	1.00	0.94	0.06
4.	I won't listen (too loud)	(27)	0.98	1.00	0.95	0.05	0.80	0.94	0.56	0.38
5.	Hello (goodbye)	(18)	0.98	1.00	0.95	0.05	0.91	1.00	0.69	0.31
6.	I'm cold	(24)	0.98	1.00	0.95	0.05	0.84	1.00	0.56	0.44
7.	I'm happy*	(36)	0.66	0.87	0.59	0.28	0.29	0.25	0.31	-0.06
8.	Tastes good	(28)	0.76	0.87	0.68	0.18	0.46	0.63	0.31	0.31
9.	Peace*	(31)	0.49	0.80	0.18	0.62	0.14	0.19	0.19	0.00
10.	What time is it?	(33)	0.90	1.00	0.82	0.18	0.86	1.00	0.75	0.25
11.	Tastes awful*	(26)	0.66	1.00	0.41	0.59	0.63	0.69	0.69	0.00
12.	Get up	(25)	0.51	0.80	0.27	0.53	0.75	0.88	0.50	0.38
13.	You	(5)	0.59	0.87	0.36	0.50	0.86	0.94	0.69	0.25
14.	Blowing a kiss love	(20)	0.63	1.00	0.32	0.68	0.71	0.88	0.50	0.38
15.	No Good	(4)	0.85	1.00	0.73	0.27	0.27	0.50	0.00	0.50
16.	Over there	(37)	0.78	0.87	0.77	0.09	0.39	0.50	0.25	0.25
17.	Louder (I can't hear you)	(30)	0.95	1.00	0.91	0.09	0.96	1.00	0.88	0.13
18.	I'm tired	(21)	0.37	0.73	0.14	0.60	0.00	0.00	0.00	0.00
19.	Crazy	(13)	0.93	1.00	0.86	0.14	0.89	0.94	0.69	0.25
20.	I don't know	(2)	0.87	0.93	0.82	0.12	0.71	0.94	0.38	0.56
21.	I'm hot	(32)	0.66	0.80	0.64	0.16	0.63	0.88	0.25	0.63
22.	Give it to me*	(11)	0.40	0.67	0.18	0.48	0.14	0.19	0.13	0.06
23.	Stop	(12)	0.93	1.00	0.86	0.14	0.89	1.00	0.81	0.19
24.	No	(15)	0.98	1.00	0.95	0.05	0.77	1.00	0.44	0.56
25.	I'm sad*	(34)	0.63	0.73	0.59	0.14	0.36	0.31	0.38	-0.06
26.	Naughty child (don't do that)*	(6)	0.49	0.93	0.23	0.71	0.05	0.06	0.00	0.06
27.	Come here	(14)	0.93	1.00	0.86	0.14	0.77	0.94	0.56	0.38
28.	Oh, no!	(3)	0.72	0.80	0.68	0.12	0.71	1.00	0.56	0.44
29.	Money*	(9)	0.67	0.93	0.45	0.48	0.23	0.25	0.19	0.06
30.	It smells bad	(29)	0.98	1.00	0.95	0.05	0.98	1.00	0.94	0.06
	Going my way (hitchhiking)	(8)	0.23	0.33	0.09	0.24	----	----	----	----
	Go away	(10)	0.38	0.33	0.41	-0.08	----	----	----	----
	Quiet	(16)	1.00	1.00	1.00	0.00	----	----	----	----
	Going to sleep	(17)	1.00	1.00	1.00	0.00	----	----	----	----
	Punch in the nose anger	(22)	0.37	0.47	0.32	0.15	----	----	----	----
	Get out	(23)	0.80	0.80	0.77	0.03	----	----	----	----
	Sit down	(35)	0.54	0.40	0.64	-0.24	----	----	----	----
	Big and round	(38)	0.66	0.67	0.68	-0.02	----	----	----	----

* Items eliminated for revised 23-item gesture test.

BEST COPY AVAILABLE

ministered to three groups of Japanese university students at their university's language laboratory. Students were able to view the video on small monitors on their desks as well as on four large televisions suspended from the classroom ceiling.

After correcting the tests, the IF and ID were calculated for each item. In this case, ID was the difference between the IF for the students who scored in the upper quarter of the 56 students for the whole test minus those who scored in the lower quarter. Descriptive statistics, including measures of reliability, were also calculated. Seven items which did not work well according to their IF and ID statistics were then eliminated and the descriptive statistics recalculated for a "revised" version of the Gestest with fewer items and improved reliability.

Results

Table 2 shows the descriptive statistics for NSs and NNSs for the pilot video with interrater reliability estimates (*r*) and for the pilot Gestest and revised Gestest with Cronbach's alpha (α) measuring internal consistency. All results appear to be normally distributed. Thanks to the careful selection of gestures and the thorough piloting of the video, the pilot Gestest already has a fairly acceptable level of internal consistency.

Table 3 shows the item analysis of the results of the pilot video and the pilot Gestest. The eight items at the bottom of the list were eliminated for the pilot Gestest. Those that had low IFs even for NSs may have had some problem with the video performance itself. This is very important to keep in mind for anyone attempting to construct a similar video test of nonverbal sociolinguistic ability.

Gestures marked with an asterisk were eliminated because they were either too difficult, too easy, or did not discriminate well enough between the highest and lowest scorers. The remaining items were used to recalculate the results as if this were a 23–item revised Gestest. As seen in Table 1, the acceptable reliability of the improved version ($\alpha = .75$) was achieved by merely removing items that did not work well.

Discussion

The above description showed how traditional norm–referenced item analysis using IF and ID can be applied to the construction of a test of nonverbal sociolinguistic ability. It was found that:

1. There are 38 gestures that NSs feel are important for language learners in Japan.
2. Thirty of these gestures can be used to discriminate between NSs and NNSs understanding of North American gestures.
3. By eliminating seven of the thirty gesture items (because of their item statistics), a leaner and more effective version of the test could be created.

As for research question 4, the use of IF and ID to construct this test after the careful collection of gesture data from NSs and NNS, clearly resulted in a more reliable test. In addition, this test also appears to be a valid test of nonverbal sociolinguistic ability at least in terms of content validity for the task itself as well for the construct being measured.

In terms of task, if these gestures collected from the literature are truly emblematic, they should be interpretable without the help of verbal or contextual clues. NSs were able to accurately identify the gestures to such a degree that many of them used exactly the same language as the English cues listed in Table 3 taken directly from the literature. ID statistics comparing NSs and NNSs provide some evidence of construct validity by showing that the ability to identify these gestures differs between North Americans and Japanese. As previously stated, sociolinguistic competence involves "a sensitivity to differences in variety . . ." (Bachman, 1990, p. 95). Item statistics revealed differences in the sensitivity toward these gestures and, along with the results of the test, provided evidence for the existence of a nonverbal sociolinguistic ability. Other evidence for validity can only be collected in the future by comparing Gestest results with the results of other yet–to–be constructed tests of nonverbal ability. A further discussion of validity and tests of nonverbal ability is included at the end of this chapter.

The development of the Gestest did not end here. Items for the revised version of the test were chosen but not retested in fact. Ultimately, a test such as this requires further fine tuning through what is called item quality analysis. By examining which distractors worked and did not work and revising the item options, the overall reliability could probably be improved further.

Item 18 on the pilot test, for example, turned out to be a problematic one due to the translation of the correct answer. "I'm tired" was meant to mean tired as in sleepy. Unfortunately, both "overworked" and sleepy "tired" were included in the options. Because of the translation process, *tsukareta* or I'm tired was inadvertently designated as the correct answer, but *nemui* or I'm sleepy was also included. No one chose the "correct" answer, but item analysis showed that *nemui* was actually the correct answer, and thus, the item analysis in Table 2 uses that answer as a basis for analysis.

Another problem encountered had to do with the use of the video performances themselves. In the case of the Gestest, some of the gestures that were dropped were poor items more because of the ambiguity in the performance than because of the lack of validity of the gesture itself. It remains to be shown whether performances by a male or a different female would produce different results.

The NOVA Scales

The Nonverbal Ability (NOVA) Scales were developed to measure language learners' nonverbal textual and strategic ability in terms of head nods, gaze direction changes, and gestures. The test consists of a role play to elicit the target role play behaviors, the *NOVA Rater's Guide* (Jungheim, 1994c) containing a description of the test, rating guidelines to train raters, and a rater training video with examples of actual role play performances.

Since this test is based on a communicative task, it was created keeping in mind Brown's (1995) suggestions for avoiding complications that may arise with this type of test construction: (a) the task was clearly defined with specific instructions for the NS tester (in English) and for the NNS examinee (in Japanese); (b) the task was narrow enough in scope to fit into the allotted time confirmed by piloting the role play task; (c) scoring procedures were worked out clearly in advance and described in the above–mentioned rater's manual; and (d) points on the rating scales were clearly defined to facilitate the raters' task; and (e) outside raters were used to make the rating as anonymous as possible.

This section will present a general outline of the steps taken to develop the NOVA Scales. The following research questions guided this approach to the construction of a test of nonverbal textual and strategic competence:

1. What nonverbal behaviors are used by language learners in conversations?
2. What kind of scales can be used to rate language learners' nonverbal ability in relation to these behaviors?
3. What are the characteristics of language learners' scores when rated using these scales?
4. To what extent are the these scales reliable and valid for assessing nonverbal ability?

Method

Subjects. The subjects in the NOVA Scales development process were 28 nonnative speakers and 20 native speakers of English. The NNSs were educated middle-class Japanese (23 males and five females) EFL learners comprising 24 students, two faculty members, and two office staff members of a Japanese university. The NSs were 20 educated white middle-class North Americans (15 males and five females) who were EFL teachers or graduate students in TESOL at a Japanese

Table 4. *Nonverbal ability (NOVA) scales*

	Textual Ability
Rating	*Frequency*
0	Extremely limited use of head nods and infrequent changes in gaze direction toward partner in conversation
1	Frequent use of head nods and changes in gaze direction that are not acceptable by native speaker norms
2	Frequency of head nods and changes in gaze direction approaches native speaker norms
3	Frequency of head nods and changes in gaze direction acceptable by native speaker norms
Appropriateness	
0	Totally inappropriate use of head nods and gaze direction by native speaker norms
1	Frequent inappropriate use of head nods and changes in gaze direction
2	Few inappropriate uses of head nods and changes in gaze direction
3	Use of head nods and changes in gaze direction acceptable by native speaker norms
	Strategic Ability
Rating	*Compensatory Usage*
0	No evidence of hand gestures to solve considerable linguistic problems
1	Limited use of hand gestures to solve linguistic problems with occasionally unsuccessful results
2	Hand gestures successfully used to solve linguistic problems
3	Few linguistic problems requiring the use of hand gestures for compensation
	Appropriateness
0	Never uses hand gestures to support or enhance meaning
1	Occasionally uses hand gestures to support or enhance meaning, often inappropriately by native speaker norms
2	Most hand gestures approach native speaker norms
3	Use of hand gestures appropriate by native speaker norms

branch of an American university in Japan. NS subjects provided the baseline data for nonverbal behaviors.

Materials. Materials used were role play cards describing the role play tasks, video tapes of the subjects performing the role plays, the *NOVA Rater's Guide* for training raters, training videos, and rating sheets.

Procedures. NNS subjects were recruited through either a poster placed on a university bulletin board or personally by the researcher. Next, they were paired according to their English proficiency level using previously administered linguistic and oral proficiency tests. This was followed by their performance of a thoroughly piloted series of four role plays (described in detail in Jungheim, 1994a). The first three were performed by each pair. The fourth (see Appendix B for role play instructions), which was used for the application of the NOVA Scales, was performed individually with the researcher's NS assistant. The second and third were then repeated in Japanese. Subjects performed the role plays in front of two video cameras in the researcher's office.

To collect baseline data, NSs followed the same procedure for performing the role plays in rooms at their university, with the exception of the repetition of the second and third role plays.

Videos of the role play performances were observed and transcribed. Head nods, gaze direction changes, and gestures were chosen for analysis and coded on the transcripts of the role plays. (Details of the further analyses of the role play performances as well as the proficiency tests can be found in Jungheim, 1994d.)

As a result of the analyses of nonverbal behavior used in the role play performances, four four–point (0–3) NOVA Scales were created on the basis of the above–mentioned CLA nonverbal ability framework (Jungheim, 1994d). They included the rating of frequency and appropriateness for nonverbal textual ability and compensatory usage and appropriateness for nonverbal strategic ability. See Table 4 for the details of these scales.

The *Nova Rater's Guide* was then created to train raters. It included the role play instructions for role play four (a student requesting a letter of recommendation from his professor) used for the application of the rating scales, training instructions, the NOVA Scales, ratings of performances on the training video, explanations of the ratings to help raters interpret "native speaker norms," and a sample rating sheet. "Native speaker norms" were established by observing and analyzing the role play performances of the 20 North American NSs. This norm was chosen because of the need for a particular norm to act as a guide for rating. Since many English teachers in Japan are North Americans, their nonverbal behavior is actually an input variety for Japanese learners, and as such, it can serve as a norm (Kasper, 1992). It must be emphasized here that other norms would be equally valid, but for research purposes the North American norm was chosen.

A training video was created using pilot performances of role play four. A rating video was also created by copying the NNS subjects' performances of role play four in random order onto a separate video with each performance preceded by the subject's number.

Two raters were then given the training materials and rating video and asked to rate the performances. A follow–up rating was performed by the researcher as a third rater a month later because a number of ratings differed by more than one point between the first two raters. This was carried out to improve the reliability of the ratings. Although it is normally not desirable for the researcher to do this because of the issue of anonymity, it was justified here on the basis of the three–month lapse since he had last viewed the videos.

Finally, descriptive statistics including reliability estimates using intraclass correlation

Table 5. *Descriptive statistics for the NOVA scales*

		Textual Ability		Strategic Ability	
	Overall Score	1*	2	3	4
Mean	1.57	1.61	1.37	1.42	1.49
S	.53	.50	.85	.81	.63
Median	1.67	1.67	1.67	1.67	1.54
Low	.33	.67	.00	.00	.25
High	2.67	3.00	3.00	3.00	2.92
Range	2.34	2.33	3.00	3.00	2.67
r_{kk}	.67	.70	.81	.81	.93
SEM	.30	.29	.37	.35	.17

* 1 = Frequency, 2 = Appropriateness, 3 = Compensatory Usage, 4 = Appropriateness

for three raters were calculated for each of the four scales and the average score for the scales of the NOVA Scales.

Results

Table 5 shows the descriptive statistics for the three raters' ratings of role play four performances of the 28 Japanese subjects using the NOVA Scales. The high intraclass correlation (r_{kk}) for the test average means that there was a high degree of agreement among the raters for the test as a whole (i.e., $r_{kk} = .93$). Reliability estimates for each scale, $r_{kk} = .67–.81$, were also within acceptable limits. Textual ability ratings for the use of head nods and gaze direction changes appear to be normally distributed. The slightly positive skew (i.e. the average for the whole group is much lower than the median or "middle" score) for nonverbal strategic ability ratings may be related to the number of subjects who did not gesture at all even though they could have used gestures to compensate for their linguistic difficulties.

Discussion

The NOVA Scale results have shown that it is possible to construct a productive test of nonverbal textual and strategic ability by collecting baseline data to establish norms for the nonverbal behaviors to be tested and by taking necessary test construction precautions for creating a test based on a communicative task. It has found that:

1. Head nods, gaze direction changes, and gestures are nonverbal behaviors used by language learners in conversations.
2. Based on a CLA nonverbal ability model, nonverbal ability scales can be constructed to assess the frequency and appropriateness of the use of head nods and gaze direction changes as a part of textual ability and the compensatory usage and appropriateness of gestures as a part of strategic ability.
3. Ratings using the NOVA scales generally produce a normal distribution for textual ability but a slightly positively skewed distribution for strategic ability.
4. The NOVA Scales as a whole are a very reliable measure of nonverbal ability related to the use of head nods, gaze direction changes, and gestures for this group of students.
5. Content validity was assured by the use of the role play task, which was appropriate for the university subjects who might actually need to ask for a letter or recommendation in real life, and by reviewing the literature on nonverbal behavior and analyzing the behavior of the baseline data subjects.

The issues of reliability and validity will be discussed further in the next section.

Conclusions

This chapter has described how tests of various types of nonverbal behavior can be constructed under a communicative competence framework using standard language testing practices. However, these tests have been constructed and used primarily in a research context and are certainly not the only alternatives for assessing nonverbal behavior.

Outside of actual assessment for the purpose of grading students, such instruments can also be used as classroom tools. The Gestest, for example, has proven to be useful for heightening learners' awareness of cultural differences in nonverbal communication as a kind of pretest instrument in English conversation classes and intercultural communication seminars. Constructing it as a multiple–choice test with all of the options in Japanese has made it possible to administer the Gestest in only about 15 minutes, short enough to allow ample time for follow–up discussion in the classroom.

Problems

While the Gestest and the NOVA Scales have been shown to be reliable tests of nonverbal abilities, the issue of validity has still not been thoroughly investigated. In the above descriptions of these tests, only content validity has been dealt with. It should be noted, however, that although we often divide validity into content, criterion–related, and construct validities, validity is actually a unitary concept. Furthermore, not only is the test itself being validated but inferences are also being made about the uses of the test (APA, 1985). That is why it was emphasized that, in this study, the tests were primarily used as research instruments. The validity of these tests for other purposes would have to be considered in light of who the learners are as well as the purpose for testing their nonverbal ability.

Evidence for content validity has been provided for both tests in terms of the target nonverbal behaviors and the tasks themselves. In order to provide a fuller picture of these tests' validity, it is necessary to discuss criterion–related and construct validity.

Criterion–related validity can only be determined by comparing the results of a test with another measure of the same construct. Due to the lack of other measures of nonverbal ability, this is not possible. One answer, then, is to create other measures of nonverbal ability following the same careful steps as described for the Gestest and the NOVA Scales.

The PONS Test (Rosenthal, Hall, Archer, DiMatteo, & Rogers, 1979) offers some potential as a test of sociolinguistic competence, although its reliability and validity for language learners must first be investigated. In addition, Jungheim(1992) suggested two other tests of nonverbal ability, a gesture appropriateness test and a bicultural gesture measure. The gesture appropriateness test would consist of brief, randomly–ordered video conversations. There would be both appropriate and inappropriate uses of the target gestures in addition to performances with six other easy gestures used appropriately. Subjects would be required to decide only whether the gesture was used correctly or incorrectly. A bicultural gesture measure could be patterned after the *Bilingual Syntax Measure* (Burt, Dulay, & Hernandez–Chavez, 1975) and used as a productive measure of nonverbal ability. Subjects would perform a dialogue that required the use of certain target gestures. Raters would rate videos of the performances for the number of correct usages of gestures in obligatory occasions. Furthermore, an alternate version of the Gestest itself could also be created using the same items with different video performances and arranged in different orders. These are just some of the possibilities for creating tests of nonverbal ability that could be used to establish criterion–related validity.

Construct validity is established by showing experimentally that a test actually assesses a particular psychological construct that cannot be measured directly. Factor analysis has shown that the NOVA Scales do measure a nonverbal ability (Jungheim, 1994d), thus confirming their construct validity. Further research, however, will need to include additional measures of nonverbal ability such as the Gestest in this analysis.

Future Test Development

This chapter has provided a framework for dealing with nonverbal ability in a communicative competence context and a number of examples of tests that could be used to assess learners in this respect. Others who are interested in developing their own tests should keep in mind three important aspects of developing tests of communicative competence: "what to look for, how to gather relevant information, and how to use the information" (Canale, 1988, p. 67).

What to look for should include a thorough review of what nonverbal behaviors are important for target learners. This may vary from culture to culture. Instruments developed here were constructed with Japanese university students in mind. *How to gather relevant information* will depend on the nature of the test itself. Aside from referring to the nonverbal communication literature, it is important to establish some norm and collect at least a limited amount of baseline data. *How to use evaluation information* is ultimately the most important aspect because it involves the issue of validity. The tests described here may be valid for research or as a pedagogical tool but not necessarily to evaluate the nonverbal ability of hotel trainees or teaching assistants at an American university. Teachers will need to take a hard look at how the results of their tests will be used.

It is hoped that this chapter will stimulate interest in the assessment of nonverbal ability and further contribute to the refinement of a framework for testing as well as teaching the use of nonverbal communication by foreign language learners.

Note

Anyone who is interested in obtaining a copy of the *NOVA Rater's Guide* or further information about any of the other materials described here should write the author at 5–22–10 Shimoigusa, Suginami–ku, Tokyo 167, Japan.

References

Al–shabbi, A. E. (1993). Gestures in the communicative language teaching classroom. *TESOL Journal, 2*(3), 16–19.

APA. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze.* Cambridge, England: Cambridge University Press.

Bachman, L. S. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition, 10,* 149-163.

Bachman, L. S. (1990). *Fundamental considerations in language Testing.* Oxford: Oxford University Press.

Bachman, L. S. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671-704.

Bachman, L. S., & Palmer, A. S. (1983). *Oral interview test of communicative proficiency in English.* Urbana, Ill.: Photo-offset.

Bachman, L. S., & Savignon, S. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The Modern Language Journal, 70*(4), 380-390.

Baird, L. L. (1983). The search for communication skills. *Educational Testing Service Research Report,* No. 83-14.

Barnlund, D. (1989). *Communicative styles of Japanese and Americans: Images and realities.* Belmont, CA: Wadsworth.

Bialystok, E. (1990). *Communication strategies: A psychological analysis of second-language use.* Oxford: Basil Blackwell.

Brown, J. D. (1995). *Testing in language programs.* Englewood Cliffs, NJ: Prentice–Hall.

Burt, M. K., Dulay, H. C., & Hernandez–Chavez, E. (1975). Bilingual *syntax measure I.* San Antonio, TX: Psychological Corporation.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. Richards & R. Schmidt (Eds.), *Language and communication.* London: Longman.

Canale, M. (1988). The measurement of communicative competence. *Annual Review of Applied Linguistics, 8,* 67-84.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second-language teaching and testing. *Applied Linguistics, 1,* 1-47.

Duncan, S., Jr., & Fiske, D. W. (1985). *Interaction structure and strategy.* Cambridge: Cambridge University Press.

Ekman, P. (1976). Movements with precise meanings. *Journal of Communication, 3,* 14–26.

Greenbaum, P. E. (1985). Nonverbal differences in communication style between American Indian and Anglo elementary classrooms. *American Educational Research Journal, 2*(1), 101–115.

Hattori, T. (1986). A study of nonverbal intercultural communication between Japanese and Americans focusing on the use of the eyes. *JALT Journal, 8,* 109–118.

Hurley, D. S. (1992). Issues in teaching pragmatics, prosody, and non–verbal communication. *Applied Linguistics, 13*(3), 259–281.

Johnson, H. G., Ekman, P., & Friesen, W. V. (1975). Communicative body movements: American emblems. *Semiotica, 15*(4), 335–353.

Jungheim, N. O. (1991). A Study on the classroom acquisition of gestures in Japan. *The Journal of Ryutsu Keizai University, 27*(2), 61–68.

Jungheim, N. O. (1992). Interactions of explicit and implicit knowledge and the effect of practice on the acquisition of L2 gestures. Unpublished paper, Tokyo: Temple University Japan.

Jungheim, N. O. (1994a). *Assessing the nonverbal ability of foreign language learners.* Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, La, April 4–8, 1994). (ERIC Document Reproduction Service No. 374 676).

Jungheim, N. O. (1994b). Through the learner's eyes: Nonverbal behavior and personality in the foreign language classroom. *Temple University Japan Research Studies in TESOL, 2*, 93–108.

Jungheim, N. O. (1994c). *Nonverbal ability assessment scales: NOVA rater's guide*. Unpublished manuscript, Temple University, Tokyo.

Jungheim, N. O. (1994d). *Assessing nonverbal ability as a component of language learners' communicative competence*. Unpublished doctoral dissertation, Temple University, Tokyo.

Jungheim, N., & Ushimaru, A. (1990). Kaiwa tatsujin e no michi "pafomansu" ("Performance"—A key to the art of conversation). *Hyakumannin no Eigo* (English for Millions), *5*, 20–25.

Kasper, G. (1992). Pragmatic transfer. *Second Language Research 8*(3), 203-231.

Kellerman, S. (1992). 'I see what you mean': The role of kinesic behaviour in listening and implications for foreign and second language learning. *Applied Linguistics, 13*(3), 239–258.

Kimura, T. (1979). Language is culture: Culture–based differences in Japanese and English. In H. M. Taylor (Ed.) *English & Japanese in contrast*. New York: Regents.

Kumin, L., & Lazar, M. (1974), Gestural communication in preschool children. *Perceptual and Motor Skills, 38*, 708–710.

La France, M., & Mayo, C. (1978). Gaze direction in interracialdyadic communication. *Ethnicity, 5*, 167-73.

Maynard, S. K. (1987). Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation. *Journal of Pragmatics, 11*, 589-606.

Maynard, S. K. (1989). *Japanese conversation: Self-contextualization through structure and interactional management*. Norwood, NJ: Ablex.

Maynard, S. K. (1990). Conversation management in contrast: Listener response in Japanese and American English. *Journal of Pragmatics, 14*, 397-412.

Mohan, B., & Helmer, S. (1988). Context and second language development: Preschoolers' comprehension of gestures. *Applied Linguistics, 9*(3), 275–292.

Nambiar, M. K., & Goon, C. (1993). Assessment of oral skills: A comparison of scores obtained through audio recordings to those obtained through face-to-face interaction. *RELC Journal, 24*(1), 15-31.

Neu, J. (1990). Assessing the role of nonverbal communication in the acquisition of communicative competence in L2. In R. C. Scarcella, E. S. Andersen, & S. D. Krashen (Eds.). *Developing communicative competence in a second language*. New York: Newbury House.

Pennycook, A. (1985). Actions speak louder than words: Paralanguage, communication, and education. *TESOL Quarterly, 19*(2), 259–282.

Rosenthal, R., Hall, J. A., Archer, D., DiMatteo, M. R., & Rogers, P. L. (1979). *The PONS test manual: Profile of nonverbal sensitivity*. New York: Irvington Publishers.

Rubin, R. B. (1982). Assessing speaking and listening competence at the college level: The communication competency assessment instrument. *Communication Education, 31*, 19-32.

von Raffler-Engel, W. (1980). Kinesics and paralinguistics: A neglected factor in second-language research and teaching. *Canadian Modern Language Review, 36*(2), 225-37.

Appendix A: Gestest Answer Sheet

学生番号		氏名		年令	才	性別	男・女

ビデオの人物が何かを表現しようとしています。下記の中から最も ふさわしい答えを選び適切な記号を塗りつぶしなさい。

1. a. お金を貸して！ □a □b □c □d
   b. オーケー！
   c. わかった！
   d. なかなか！

2. a. 私ですよ。 □a □b □c □d
   b. ハートよ。
   c. 胸焼けだ。
   d. ここです。

3. a. 気持ちいいんだ。 □a □b □c □d
   b. 肩が凝っている。
   c. 早く言いなさい。
   d. はい、わかった。

4. a. はずかしい。 □a □b □c □d
   b. 耳が痛い。
   c. うるさいのよ。
   d. あー、心配。

5. a. 止めなさい。 □a □b □c □d
   b. もうないよ。
   c. さようなら。
   c. ねー、聞いて。

6. a. 恐かった。 □a □b □c □d
   b. 抱いてくれ。
   c. さむーい。
   d. 体が痛い。

7. a. いいですね。 □a □b □c □d
   b. わからない。
   c. おもしろい。
   d. うれしい。

8. a. おいしい。 □a □b □c □d
   b. べろべろばー。
   c. へーんだ。
   d. のどが乾いた。

9. a. 大丈夫だ。 □a □b □c □d
   b. やった。
   c. 平和よ。
   d. 勝った。

10. a. 待てない。 □a □b □c □d
    b. 遅れている。
    c. 今、何時。
    d. 時計はどこ。

11. a. アカンベー。 □a □b □c □d
    b. ひどいよ。
    c. まずいよ。
    d. いやーよ。

12. a. もりあがって。 □a □b □c □d
    b. 立ちなさい。
    c. うえ、うえ。
    d. 早く、早く。

13. a. あなたよ。 □a □b □c □d
    b. そこだ。
    c. あそこ見ろ。
    d. 来なさい。

14. a. どうぞ。 □a □b □c □d
    b. 好きよ。
    c. あまい。
    d. またね。

15. a. ダメー！ □a □b □c □d
    b. ちょっと！
    c. さげろ！
    d. この野郎！

16. a. 違うよ。 □a □b □c □d
    b. 出なさい。
    c. あの人。
    d. 向こうよ。

17. a. よく聞いてくれ。 □a □b □c □d
    b. エッ？聞こえない。
    c. うるさいわね。
    d. わからないのか。

18. a. あきたよ。 □a □b □c □d
    b. ねむーい。
    c. つかれた。
    d. つまらない。

19. a. 頭がかゆい。 □a □b □c □d
    b. 頭がおかしい。
    c. 頭が痛いの。
    d. 頭にきた。

20. a. いまいちだ。 □a □b □c □d
    b. 情けない。
    c. わからない。
    d. 困ったな。

21. a. なかなか。 □a □b □c □d
    b. あー、暑い。
    c. 危なかった。
    d. 疲れたよ。

22. a. ちょうだい。 □a □b □c □d
    b. お入りなさい。
    c. 見てください。
    d. どうぞ。

BEST COPY AVAILABLE

23. a. あっちに行け。 a b c d
    b. まだまだ。
    c. お手上げ。
    d. 止まれ！

24. a. 違います。 a b c d
    b. わからない。
    c. 残念です。
    d. 目が回る。

25. a. 困ったなー。 a b c d
    b. 悲しいわ。
    c. しらなーい。
    d. 失敗した。

26. a. 悪い子！ a b c d
    b. いいですか。
    c. 違います。
    d. そのとおり。

27. a. 見せて下さい。 a b c d
    b. 貸しなさい。
    c. こっちへ来て
    d. 見てごらん。

28. a. 頭が痛い。 a b c d
    b. どうしよう！
    c. まいった。
    d. 驚いた。

29. a. 少しだけ。 a b c d
    b. 汚いな。
    c. お金よ。
    d. もう少し。

30. a. かゆーい。 a b c d
    b. くさーい。
    c. 鼻がでる。
    d. 鼻がつまった。

Appendix B
Role Play Instructions for Nonnative Speakers

You will be asked to perform a number of role plays. Read each situation and imagine that you are really the person in the role play. Specific details are not given, but try to speak in as much as detail as possible.

Nonnative Speaker Subject's Role

You want to apply to an American university's graduate school to study in your major. You need a letter of recommendation from your teacher. You go to your teacher's office to ask him for one. Be ready to explain why you need the letter, what information it should contain, when you need it, and what your teacher should do with it in as much detail as possible.

Instructions for Native Speaker Assistant

1. Greet the student: "Hi, (name), what can I do for you?"
2. Remember. You do not know what the student wants beforehand.
3. Help the student speak, but do not put words in his/her mouth. The burden to communicate should be on the student as long as pauses are not too long. Be patient.
4. If the student does not give instructions about the letter of recommendation, ask questions about why he/she wants to study abroad, the information to be contained in the letter, when it is needed, where to send it, and so on.
5. You will have four minutes for the role play, but do not worry about the time. If there is time, close the conversation with small talk and a word of "good luck."

Native Speaker Assistant's Role

You are a professor at a Japanese university. A student comes to you to ask for a letter of recommendation so that he can study at an American university. Ask the student why he wants to study abroad, what information the letter should contain, when he needs it, and where he should send it. Try to get the student to explain in as much detail as possible.

Chapter 17

Cloze Testing Options for the Classroom

CECILIA B. IKEGUCHI
DOKKYO UNIVERSITY

Controversies in the field of cloze testing research are far from solved, and the debate on the pros and cons of cloze procedure continues. Whatever has been said and done, the merits of cloze testing seem to outweigh its demerits (Chapelle & Abraham, 1990). Although research reports come up with inconsistent results from one cloze to another, cloze appears to be a measure of distinct language skills (Bachman, 1990). Aside from these empirical inconsistencies, there exists an unfortunate reality in terms of the practical applications of the cloze. It is a fact that the concept of cloze testing has often been confined within the walls of empirical research, "R," which is often considered irrelevant and far-out from actual classroom testing situations, "r" (Lo Castro, 1994). This chapter aims to bring research on cloze testing into the classroom setting by introducing simplified ways of testing language skills by using cloze. Specifically, this chapter aims first, to provide a simplified review of the research on cloze testing for practical classroom uses; it does not attempt to be comprehensive. In so doing, I wish to orient the language teacher to what's going on in the research related to their actual teaching practices. This chapter seeks as much as possible to avoid theoretical discussions (since they are given in voluminous publications elsewhere), and the terminology is simplified. If theoretical background information is included, it is for the purpose of providing the necessary theoretical justifications for the practicality of cloze in the classroom.

Second, this chapter seeks to highlight the implications of empirical findings for classroom testing. More specifically and most importantly, it aims to present a systematic discussion of the four types of cloze tests (the fixed-rate deletion, the rational, the multiple–choice cloze, and the C-test), their theoretical justifications, the features of each, as well as their distinct merits and demerits. Since the accuracy of measuring what the language teacher wants to assess may depend on the kind of deletion procedure that is selected, each procedure will be discussed in turn.

So much has been written about cloze testing in general that, on first thought, it seemed redundant to include a history of cloze testing in this chapter. However, to understand the different cloze deletions that are central to the present discussion, it is necessary to get a bit of information on the precedents and causes that gave birth to cloze testing.

Background of Cloze Testing

In the cloze procedure, the examinees are given a segment from which words have been deleted and replaced by blanks, and they must provide or choose the word that best fits the blank. The theory behind cloze is that a language learner, presented with a piece of language mutilated in this manner can use his or her acquired competence to restore either the word in the original text or an acceptable word. As learners develop language competence their ability to use clues from the text to restore missing items increases (Klein-Braley & Raatz, 1984).

The seeming simplicity of this technique, regardless of its theoretical justification, has made it an attractive instrument for both test constructors and classroom teachers. Cloze testing has been regarded by researchers from two opposing and yet complementary theoretical viewpoints. Some researchers claim that cloze is an integrative rather than a discrete-point test because it draws at once on the overall grammatical, semantic, and rhetorical knowledge of the language. To reconstruct the textual message, students have to understand key ideas and perceive relationships within a stretch of continuous discourse, and to produce, rather than simply recognize, an appropriate word for each blank. The focus of the task involved is more communicative than formal in nature, and is therefore considered to reflect a person's ability to function in the language (Hanania & Shikhani, 1986).

On the other end of the theoretical continuum, are those researchers who argue that cloze testing measures only basic skills, and lower levels of reading comprehension (Shanahan, Kamil, & Tobin, 1982). In answering a cloze item, the examinee relies on clues within the immediate environment of the blank, and as such, this type of test is only measuring lower order skills. Correlational experiments with cloze reveal that cloze tests are correlated more closely with grammar tests than with reading tests (Alderson, 1979).

My aim here is not to argue for or against any of these theories; I propose rather to outline those language skills, whether they be grammar, reading, or textual skills, which the four types of cloze are believed to assess. Validity research, after all, goes on and on. Meanwhile, teachers might as well make the best use of what research to date can offer.

Originally developed by Taylor in 1953 to measure text difficulty for native readers, other researchers have since shown that cloze can likewise measure foreign and second language proficiency as well (Anderson, 1976). Cloze can most appropriately be described as a learner-centered teaching and testing device in second language situations since it is thought to challenge the efficiency of the developing L2 grammar of a student in a way that reflects natural language processes (Oller, 1972). Moreover, cloze provides a contextualized challenge to learner grammar efficiency, displaying an inevitable simplicity which Oller calls nothing short of a stroke of raw genius (Jonz, 1976). This chapter maintains that cloze testing does hold potential for measuring aspects of students' written grammatical competence, "knowledge of vocabulary, morphology, syntax, and phonology," as well as textual competence, "knowledge of cohesive and rhetorical properties of text" in second language (Bachman, 1990). The specific traits measured by a particular cloze test probably depend in part on methods of test construction, each of which will be described below.

Fixed-rate Deletion Cloze

Rationale for Fixed-rate Cloze

The rationale for using the cloze procedure to measure the examinee's reading comprehension is that, if they understand the structure and content of a text, then they will be able to utilize redundancy in the text to recover the deleted words at a better-than-chance level. Cloze test scores are said to be a direct measure of a text's redundancy because they are obtained by deleting words from the text and asking the examinees to recover the deleted words. In this context, text redundancy refers to the degree to which the language in a text is predictable when only parts of it are unknown.

Originally, there were two deletion schemes for constructing cloze texts: the fixed-ratio (also called random) deletion, and the rational–deletion method. In the fixed-ratio deletion, the test writer deletes every *nth* word, and in so doing produces a semi–random sample of the words from the passage. In rational–deletion, the test writer chooses the words to be deleted in advance, e.g., content words, or pronouns, or prepositions, etc.

In fixed-rate deletion, the procedure for omitting words in a regular pattern inevitably samples various types of words, some of which are governed by local grammatical constraints, others of which are governed by long-range textual constraints. This means that the students' ability to answer the local constraint test items depends on their ability to perceive and use grammar rules and their relationships, while answering long-range constraint items requires the students' overall comprehension of the passage.

Studies have shown, however, that differences in test results, subject to text difficulty and topic, are problematic for fixed–rate cloze testing (Brown, 1983). Alderson (1983) claims that differences in cloze test results are due not to deletion frequencies but rather to differences in the particular words deleted. According to Alderson, random deletion ignores the syntactic and semantic relationships in a text, and is therefore likely to yield inconsistent results depending upon what proportion of syntactic and textual functions are tapped. Those who support the use of cloze, Brown (1991) for instance, suggest that cloze items assess a wide range of language points from morphemes and grammar rules at the clause level to pragmatic level rules of cohesion and coherence, as well as discourse levels.

Format of the Fixed–ratio Cloze

It seems to me that evidence supporting the claim that fixed–rate cloze items test different aspects of the examinee's language ability outweigh the criticisms against it, thus I will continue the discussion along this line.

The fixed-ratio cloze test is constructed by deleting words from a passage according to a fixed-pattern, traditionally every 7th word of a text. Why every 7th? Brown's (1983, 1988a) experiments revealed that cloze passages can be made to fit a group when the distance between items (or words in a passage) was no less than five words, and no more than nine, giving an average of seven words. Although the 7th–word deletion seems to be a popular deletion pattern, 7th–word cloze test results sometimes reveal that they are too difficult for particular groups of students—with say an average score of 25%. Brown (1983) suggested ways that can be used to increase the mean level and make the test less difficult such as by lengthening the passage, and increasing the distance between the blanks, from say every 7th to every 11th, or even every 15th word.

In constructing a fixed–rate cloze, in most cases, one or two sentences are left intact (i.e., no deletions are made) at the beginning of the passage, and one or two sentences are also unmodified at the end of the passage to provide a complete context. There is usually a total of 30-50 blanks to be completed. Making an answer key while constructing the test will reveal to the teacher what words are being tested and the percentage of the different kinds of words deleted or what balance of content and function words is being sampled. Usually, a 50–item cloze test is long enough to represent the different aspects of grammar, vocabulary, and comprehension skills being checked at the end of a course unit. Language teachers, especially those in the high school, usually give a term test or a course-end language check composed of 50-60 points of grammar and vocabulary. This traditional test type could easily be replaced by two fixed–rate cloze tests of at least 30 items each. University students usually take 30-40 minutes to answer a 40-50 item fixed–rate cloze. Because of lack of research, high school language teachers will have to estimate the time needed by their students to answer a fixed–rate cloze test at the appropriate level of difficulty. However, high school students may be very similar to university students; we just don't know.

Scoring is usually done in one of two ways: the exact-word scoring method, and the acceptable-word scoring method. The exact–word scoring method refers to any scoring method where only the word that was originally in the blank in the original passage is counted as correct. The exact-word scoring system has its merits for cloze tests used for research in that: (a) it yields high correlations with other scoring methods, and (b) a single correct answer for each blank is essential for some linguistic analyses.

For classroom testing, however, the acceptable–word scoring system has more advantages than disadvantages. True, the acceptable–word scoring system can be troublesome and a bit time consuming to score, but this can be overcome if teachers meet ahead of time and decide which answers are correct and acceptable. It pays to exert a little extra effort at this stage because acceptable–word scoring is easier for students accept, yields higher performance results, and often turns out to have higher test reliability.

What Skills Does Fixed–ratio Cloze Assess?

It has been argued that there is a dichotomy between language core proficiency tests of linguistic skills of a relatively low-order and higher-order skills tests. Language skills belonging to the low–order category include grammar and vocabulary, while the skills of reading comprehension might more appropriately be classified as higher–order skills. Some cloze researchers have argued that cloze items primarily assess students' ability to comprehend words and ideas at the sentence level, which are said to be lower–order skills. On the other hand, researchers claim that cloze items measure intersentential language components, which are said to be higher-order skills. If cloze assesses higher-order skills, they would probably include students' abilities to comprehend meaning and structural relationships among sentences.

Research using the fixed-rate deletion and supporting the lower–order skills hypothesis includes Shanahan, Kamil, & Tobin (1982) and Porter (1983). In the Shanahan et al. (1982) study, which used intact and scrambled versions of cloze tests, no significant difference was found in students' performance in these the two types of cloze tests. This has led researchers to conclude that cloze tests do not measure the ability to comprehend information beyond the sentence level. Porter used cloze tests varying the points at which the deletions started. His findings reveal that cloze items are not sensitive to contexts beyond five to six words.

However, I feel that the empirical findings from the other camp outweigh the arguments just posed. For example, Brown (1988b, 1991), using the fixed-rate deletion pattern, has argued that words in a cloze test provide a fair representation of the text at the word, clause, and sentence level, i.e., at least some items are sensitive to constraints ranging beyond the limits of sentence boundaries. If true, this would mean that cloze test items do assess a student's comprehension at the discourse and pragmatic level, which definitely involves higher-order skills. Chavez-Oller, Chihara, Weaver, and Oller, Jr. (1985) made similar claims that at least some items in a cloze test are sensitive to long-range constraints beyond five-ten words. Yamashita's (1994) experiment, although limited to the morphemic and clausal level, concluded that cloze tests using the fixed-rate deletion are effective in analyzing the reading comprehension of native and non-native English learners.

This discussion of both sides of this theoretical issue are not intended to confuse the language practitioner, but are instead aimed at providing some background on what is happening in cloze testing. More importantly for the language teacher, these empirical results should provide a total picture of the reality of language testing. As for the higher-order and lower-order skills issue, the concern of most language teachers is to assess some of or both of these skills during the time available.

I feel that, when teachers make cloze tests using the fixed-rate deletion, they can rest assured that both of these levels of skills are being assessed and that each of the following structural categories is proportionately represented in the cloze test they are making: (a) word level, (b) clause level, (c) sentence level, and (d) paragraph level (Perkins & German, 1985). A word level deletion, also called local level, means that for the students to retrieve the deleted word, they rely only on the clues found within two words of a blank. A clausal level deletion means that students can find clues to answer the blank from within the clause but beyond two words of where the blank appears. A sentence level deletion means that students must find clues within the sentence where the blank appears. Paragraph level deletion means clues are found within the paragraph in order to retrieve the missing word. Thus the examinee must have understood at least part of the content of the entire paragraph in order to fill in the deleted words.

Previous research where these different levels of deletion were made on a single text is reported in Perkins and German (1985), where 10, 11, 5, and 9 deletions were made at the word, clause, sentence, and paragraph level, respectively. Chapelle and Abraham (1990) following these structural categories, had 4, 12, 6, and 3 such deletions, respectively. To me, this research clearly indicates that even in a single passage, teachers can construct a cloze test that can assess different levels of skills by systematically varying the number of deletions in each of the categories listed by Perkins and German (1985).

Pros and Cons of the Fixed-rate Deletion Pattern

The fixed-rate cloze is the most difficult of all the cloze types to answer because no deliberation over items takes place during test construction. In addition, the literature on fixed-rate deletion cloze is marked by inconsistent results from one study to another. These inconsistencies, which appear to occur even when different deletion rates are used for the same text, may be caused by uneven sampling of skills measured from one cloze to another (Alderson, 1979). Another explanation given for cloze result inconsistencies are text and item difficulty, suggesting the need for passage fit (Brown, 1983). This probably means that teachers should use passages that fit their students in terms of difficulty level, topic, interest level, and the like. Obviously, cloze items are at the root of cloze test performance, and it may be possible to improve a cloze test by explicitly selecting the words to be deleted through the use of the rational-deletion cloze—the topic which I will consider next.

Rational-Deletion Cloze Technique

Rationale for Rational-deletion Cloze

While fixed-ratio cloze tests rely on regular sampling of words in a text by deleting words in a regular pattern, the rational cloze assumes that the different cloze items can be explicitly chosen to measure different language traits. Bachman (1982) provides evidence that the test writer can select words reflecting distinct aspects of the learners' grammatical and textual competence.

Considerable debate has occurred among cloze researchers as to whether all deletions in a cloze passage measure the same abilities. Some researchers had assumed that examinees' item responses to function word deletions provides information about their understanding of the written text's structure while their item responses to content words provides information about their comprehension of the written text's content. It would seem then that rational deletion cloze could be designed to measure a wide range of language abilities.

To develop a test that would potentially measure textual relationships (students' comprehension of structure and meaning) beyond the clause level, it would be necessary to identify a set of criteria for classifying and selecting the words to be deleted, perhaps using the following three deletion types: (a) syntactic (understanding of words and their relationship within clause), (b) cohesive (based on the student's understanding of meaning and structure in the intersentential level), (c) strategic (depends on long range patterns of coherence, i.e., comprehension of the passage).

Halliday and Hasan (1976) developed a framework that can be used for constructing items that assess cohesion, but such a framework proved to be difficult and subjective, and almost impossible for the classroom teacher to apply. Another attempt was made by Bachman (1985) in which four types of deletions were defined according to a four hypothesized levels of context required for closure: (a) within clause, (b) across clauses but within sentence, (c) across sentences, within text, and (d) extra-textual. To make a test primarily based on cohesion (and thus, testing higher the order skills of comprehension), the test maker would have to maximize deletions of types (b) and (c), and minimize deletions of types (a) and (d).

Skills Measured by Rational Cloze

Bachman's (1982) study found support for the claim that cloze tests can be used to measure higher order skills of cohesion and coherence if a rational deletion procedure is used. Having control over the particular words deleted, the language teacher can construct cloze passages that measure textual relationships beyond clause boundaries. The target skills in a rational cloze can include one or all of the following: (a) syntax (depending on the students' understanding of clause level), (b) context cohesive (depending on the students' comprehension of text, based on interclausal and intersentential clues of cohesion), and (c) strategic (depending on the students' comprehension of text, based on patterns of coherence).

If a teacher makes deletions using the categories just outlined, for example, a 30 item text with an average deletion of one word in twelve could probably be administered and completed in about 20 minutes by university students. Otherwise, a teacher could use the same contextual categories as those suggested in the section on fixed–rate deletion to explicitly choose items for a rational cloze test. For example, a rational cloze test with 3, 13, 5, 14 words deleted in the word, clause, sentence, and paragraph level, respectively, which was given to university students, took 25 minutes to answer. These time limits are merely meant to be suggestive; teachers may have to modify the time allowed for students to complete a particular cloze test depending on their level, and on the kind of deletion patterns made.

Scoring a teacher-made rational–deletion cloze test can be done using either the acceptable–word or alternative–word criterion method (wherein a list of syntactically or semantically alternative answers can be prepared). Ideally, the list of acceptable answers will be made based on the responses of samples of native English speakers—if pilot-testing of the cloze test can be done. Almost always, however, using native speakers to establish an answer key is quite time consuming and therefore not possible. The simplest way to make an answer key may be to draw up a list based on the judgements of the test development team or the teacher(s) making the test.

Advantages and Disadvantages of Rational–deletion Cloze

Empirical results on rational–deletion cloze item performance have been mixed. For instance in Bachman's (1985) study, the rational–deletion cloze produced a test that was easier than the fixed-ratio cloze, while other results have shown that the overall test difficulty of the two types of cloze were almost the same (Chapelle & Abraham, 1990). The rational–deletion cloze has also been shown to produce results with higher reliability than the fixed-ratio, but also, results have been produced with about the same reliability for both rational and fixed–deletion cloze tests (Bachman, 1985).

Multiple-choice Cloze Technique

Rationale for Multiple–choice Cloze

Both the fixed-ratio (or random deletion) and the rational–deletion cloze discussed above are administered to the students with instructions to replace the missing items. The multiple-choice cloze gives students a limited range of choices with which to compare the responses they generate. The self-generated response in answering an multiple–choice cloze is what Jonz (1976) called the student-centered remodeling of the cloze procedure. However, the most important characteristic of the multiple–choice cloze is its ability to assess a wide range of language skills.

Generally, research on this type of cloze indicates that the number of items can be reduced without sacrificing the reliability of the test. Thus a shorter multiple–choice cloze can reduce the strain the student is exposed to and may represent a marked improvement from the teacher or test administrator point of view because it introduces objectivity in scoring in addition to the smaller number of items to score. Research has demonstrated that constructing a test response (as in a fill-in type of cloze) is more difficult for test takers than selecting one (as in a multiple–choice cloze). Some language teachers might ask: Doesn't making a test less difficult endanger its validity and reliability? The answer is *no.* A shorter and easier test can be just as high in reliability as a longer and more difficult one. The research indicates that comparable reliability levels can be obtained for multiple–choice and fill-in versions of the cloze test, even if the multiple–choice versions are shorter.

For example, Bensousan and Ramraz (1984) show higher mean scores obtained on multiple–choice cloze tests (which are therefore easier than the other cloze types). Other studies show that multiple–choice cloze produces adequate reliabilities. For instance, Jonz (1976) found a reliability of .76; Bensousan and Ramraz (1984) report .82 and .84; and, Hale, Stansfield, Rocks, Hicks, Butler, and Oller (1989) show a reliability of .88 for multiple–choice cloze.

While, multiple-choice variants of the cloze procedure have been less thoroughly researched than others, this type of cloze seems to involve processes similar to those required by the traditional cloze types. For instance, research on multiple–choice cloze testing shows that multiple–choice cloze obtains comparable levels of correlation with criterion measures (Hale et al. 1989).

Pros and Cons of Multiple–choice Cloze

Deletion rate. The deletion process involved in a multiple–choice cloze can either be random deletion or rational deletion, (while providing a set of other responses to choose from). Whichever deletion pattern is used, research has shown the reliability of multiple–choice cloze tests to be higher than .70 (.76 in Jonz, 1976 & 1990 and .76 in Chapelle & Abraham, 1990).

Distractors. The major drawback with the use of the multiple–choice cloze lies in the fact that it is difficult to construct. Hinofotis and Snow (1980) argue that constructing a multiple–choice cloze is a considerably more complicated procedure than constructing an open-ended cloze test. The most important consideration in making multiple–choice cloze items is the process of providing distractors (i.e., the words, other than the correct answer, from which the student must choose).

Obtaining the distractors used to create a set of four options usually requires pretesting and involves checking students' responses for item facility and discrimination. The results of pretesting can be used to determine two things: the range of possible correct responses, and distractors for the final version of the test. The highest frequency acceptable response can become the correct response on the multiple–choice test, and the unacceptable responses with the highest frequency can then be chosen as distractors (Jonz, 1976). Based on item analysis of the pilot administration, items that do not discriminate well can also be discarded or modified.

A simpler procedure that can be adopted is that used by Chapelle and Abraham (1990). In most of their items, there were four alternatives given, and these distractors were the same part of speech as the correct answer. The results yielded high reliability.

Number of test items. A short (say, thirty items) multiple–choice cloze has the following advantages: it takes the slowest university student 20 minutes to complete, and it takes only a minute to score each paper. In addition, such a test yields high reliability (Jonz, 1976), which indicates that the quality of multiple–choice cloze is on par with other cloze test types. There is no minimum or maximum number of items required to come up with a good multiple–choice test, but apparently, one-third to one-tenth fewer items can be used on multiple–choice cloze as compared to the other formats without sacrificing reliability.

Test administration. It has also been argued that any multiple–choice cloze which calls for a separate answer sheet (either containing lists of multiple-choice options or a machine response form) tends to distract the examinee and is less desirable than the fill-in cloze (Jonz, 1976). Generally, for classroom use, no separate sheet will be necessary because the options can simply be aligned in a rectangular frame within the text. A separate answer sheet is only really needed in order to facilitate scoring procedures in large scale testing. In a classroom situation, where immediate feedback be desirable, checking could even be done by the students with the teacher's guidance.

What Skills Does Multiple–choice Cloze Measure?

Now to the most important section which I believe every concerned language teacher is curious about: which area of the second language competence does an multiple–choice test assess? The section above mentioned that the deletion process chosen by the teacher in constructing a multiple–choice test will depend on what skills are targeted for the particular test. There are several empirically supported ways of constructing multiple–choice cloze.

Generally, multiple–choice test items require the reader to focus on a specific amount of text in order to answer a question. Such focus can vary from one word within the sentence to the entire text. The ability to answer multiple–choice cloze items is not restricted to the comprehension of single words; it is likewise believed to tap understanding of a wider context. Thus, it is possible to employ rational–deletion procedures to measure skills of long-range comprehension by choosing items sensitive to long-range constraints (Bensousan & Ramraz, 1984).

One study comparing multiple–choice cloze results with other cloze types indicated that multiple–choice cloze scores correlate well with composition scores (Jonz, 1976); while another study (Hinofotis & Snow, 1980) claimed that multiple–choice cloze test results correlate highest with structure and reading test scores. Most empirical findings on multiple–choice cloze show strong correlations with reading tests, supporting the assertion that multiple–choice test items can assess reading skills in ESL (Ozete, 1977; Brown, 1985; and Chapelle & Abraham, 1990). It has also been argued that the multiple-choice cloze falls closer to the discrete-point end of the test continuum.

However, I feel that the degree to which multiple–choice cloze items are discrete–point must be determined on the basis of the kinds of items actually used in the test. Items that require learners to process the meaning of connected pieces of language rather than discrete segments are integrative (Chapelle, 1988). Some multiple–choice cloze items are aimed at reading comprehension (defined in terms of textual constraints ranging across clauses) as contrasted with knowledge of grammar (short-range surface syntax, and morphology or vocabulary). A number of research studies have indicated that multiple–choice cloze can be used for teaching and testing reading comprehension. Brown (1985), for instance, recommended ways of using cloze for the teaching of reading.

The classification scheme developed by Hale et al. (1989) is the most detailed so far, and can serve as a useful reference for teachers who would like to improve their multiple–choice cloze tests. Their scheme consisted of four categories of multiple–choice cloze items, based on the three skill areas of grammar, vocabulary, and reading comprehension. In answering each cloze item, the student has to use two of these skills simultaneously, with reading always involved. Each of the four categories will be explained in turn.

Reading comprehension and grammar items. In this type of multiple–choice cloze item, the student has to understand propositional information at an inter-clausal level, emphasizing knowledge of syntax. For instance: A ballad is a folk song; however a folk song is not a ballad [because, if, whether, unless] it tells a story.

Reading comprehension and vocabulary items. To answer this type of item, the student has to comprehend inter-clausal relationships, and at the same time, knowledge of vocabulary is involved. For example: ... known as the Lost Sea. It is listed in the *Guinness Book of World Records* as the world's largest underground [water, body, lake, cave].

Grammar/reading comprehension. Items of this type tap the students' knowledge of surface syntax, while reading comprehension is involved only to the degree that the reader must understand within clause propositional information. For instance: It is generally understood that a ballad is a song that tells a story (but) a folk song is not so [easy, easily, ease, easier] defined.

Vocabulary/reading comprehension. This item type checks on the student's vocabulary skill, although it involves reading comprehension based on information within clause boundaries. For example: In fact, there are many folk songs about occupations— railroading, [following, mustering, concentrating, herding] cattle, and so on.

These examples illustrate that, although reading comprehension of a text is basic in answering a multiple–choice cloze test item, other skills such as vocabulary and grammar can also be effectively tested. Classroom teachers have control over which aspects of grammar and vocabulary they would like to emphasize depending on which of these skills is the focal point of their assessment at the time.

C-Test

Rationale for C–Test

Research on the C–test has provided ample theoretical justification and proven the empirical reliability and validity of this cloze variant. Two major investigations, Alderson (1979) and Klein-Braley (1983), pointed to a number of problems with the ordinary cloze as follows:

1. The results are unpredictable for various deletion rates. Also different deletion rates and starting points of deletion even in the same passage can cause considerable variation in difficulty.
2. Particularly for homogenous samples, cloze tests can result in unsatisfactory test reliability.
3. Students may perform differently on different cloze tests depending on the topic and difficulty level of the passage, and these variables can also result in different degrees of test reliability and validity.
4. The generally accepted practice of using only one passage for a cloze test was found to be a source of bias in scores.
5. Deleting every nth word in a cloze test may not always produce a set of words that represents the word classes found in the passage used.
6. There is the problem of lack of criterion referencing native–speaker performance. In cloze testing with L1 examinees, even adult speakers rarely obtain a perfect score, and many experience a certain amount of frustration.

In response to these problems, an alternative testing system, developed and christened by Klein-Braley and Raatz (1985) as the C–test, was developed and investigated. The C–test improves on the psychometric properties of the cloze by using an every-other-half–word deletion called the Rule of Two. First, by using several short texts from different passages in a C-test, students find it less frustrating than other types of cloze (Mochizuki, 1994) in terms of difficulty level and text topic, and thereby text bias in scores is avoided. Because of its every-other-word partial deletion pattern, the C-test was felt to adequately include words that are of the same part of speech or word class as in the texts used (Klein-Braley, 1985). The problem of lack of criterion referencing to native–speaker performance is also apparently overcome because, in experiments using the C–test, adult educated examinees have been found to achieve virtually perfect scores.

Format of the C-test

In the C-test, the second half of every second word is deleted instead of the whole word, according to the Rule of Two. A C-test consists of a number of short texts (usually five or six). The first sentence of each passage is left intact to provide context. Then beginning in the second sentence, the second half of every second word is deleted until the desired number of mutilations is reached. Each short text (usually a paragraph which consists of a series of blanks) makes up one superitem of the test, with a total of five or six superitems. The students are required to fill in the blanks with even and odd numbers of letters alternately. For example: (1) sto__ [ut], (2) ph___ [one], (3) mou__ [th], (4) ov___ [ert]. In word (1) the two letters in brackets are deleted, in word (2) three letters are deleted, in word (3) two letters, and in word (4) three letters. Numerals and proper nouns (e.g., 5100 km, Mr. James Stewart, etc.) are typically disregarded in counting every second word (Mochizuki, 1994).

To overcome the problems found in the other types of cloze tests, a list of criteria was developed by Klein-Braley and Raatz (1984) to which a C–test should comply to be satisfactory: (a) the test should have several texts: there should be at least four paragraphs each coming from different passages, (b) the test should have at least 100 deletions, (c) adult native speakers should obtain virtually perfect scores, (d) the deletions should affect a representative sample of the text, (e) only exact–word scoring should be possible, and (f) the test should have high reliability (0.80 or higher) and empirical validity.

The theoretical justifications for these rules are also given in Klein-Braley and Raatz (1984). However, busy secondary ESL teachers may not have time to comply with all of these requirements, and in fact, the results of previous empirical research may justify ignoring them.

First, according to Klein-Braley and Raatz (1984), the C–test should be constructed using 5-6 short texts. In fact, it may be more important to be concerned about the appropriateness of whatever text is used for the test in terms of reliability, validity, and of course practicality. Mochizuki (1994) reports the results of an experiment with a C-test using four different kinds of texts: Narration, Explanation, Argumentation, and Description. The results showed that C-tests based on a long passage, especially Narration and Explanation texts, were more reliable. Especially noteworthy is the fact that C-tests using the (Narration text) long passage met the requirements set by Klein-Braley & Raatz of high reliability (.90), though it had only moderate criterion–related validity (a .50 correlation with a criterion measure). Mochizuki concluded that a C-test based on a single long passage, even though there was only one kind of text, might work well with secondary school classes not only because of its high reliability results but also because of its practicality in terms of exempting the teacher from the burden of finding numerous kinds of texts.

Second, Klein-Braley and Raatz (1984) argued that a C–test should have at least 100 deletions. It is easy to come up with a 100-item test using a long narrative text or explanation passage (if it has more than 200 words) from a high school reader, a textbook passage that has not yet been taken up in class, or a supplementary text. To the teacher who wants to use a one-text test, the text should preferably be a narration or explanation passage of four or more paragraphs. Then each paragraph can be treated as a superitem with at least 15 deletions in each paragraph. In most cases, the reliability should be relatively high. Chapelle and Abraham (1990) and Ikeguchi, (unpublished ms.) each used less than 100 deletions, and yet the reliability was as high as .90. However, test makers who use C-tests for program–level decisions like placement may well want to use four or more short texts of 100 deletions each in accordance to the theoretical justifications outlined by Klein-Braley and Raatz (1984).

Third, Klein-Braley and Raatz (1984) argued that adult native speakers should achieve virtually perfect scores. I feel that ESL teachers generally need not worry about this because researchers on C-tests have shown that one of the most interesting features of the C-test is the consistently excellent empirical performance of native speakers on these tests. Educated adult native speakers almost always achieve virtually perfect scores on C-tests (Klein-Braley & Raatz, 1984).

Fourth, Klein-Braley and Raatz (1984) argued that deletions should be a representative sample of the text. I feel that the very nature of the deletion system in a C-test makes it possible for deleted words to adequately represent all of the word classes that are present in the passage. The language practitioner need not worry about this because empirical research gives considerable reason to believe that a representative sample will be created in most C–tests.

Fifth, Klein-Braley and Raatz (1984) said that only the exact-word scoring method should be possible. Typically, only the exact-word answers occur on a C-test. Hence, scoring has been shown repeatedly to be a lot easier for the C-test than for most other cloze test types. One word of advice to make scoring much easier is to have students use a separate answer sheet where they will write the whole word answers. It is easier for scorers to recognize whole words rather than recognize groups of letters. This strategy will lead to more practical, faster, and efficient scoring.

Sixth, Klein-Braley and Raatz (1984) say that a C–test should have high reliability and validity. Chapelle & Abraham (1990) used one long passage with five paragraphs for their C- test experiment. Each paragraph was treated as a superitem with fifteen deletions in each paragraph. The reliability was found to be .81, and the lowest validity coefficient was .50. In Mochizuki's (1994) experiment, one long passage of the Narration type was used, with five paragraphs having a series of deletions. The results were higher than .90 reliability and .50 criterion–related validity. The results for Ikeguchi (unpublished ms.), which used one long passage with five paragraphs containing approximately fifteen deletions each, indicated a reliability of .90 and a .60 correlation with a criterion–related validity measure.

What Does C–test Measure?

The next important question that has to be dealt with at this point is what specific language traits does the C-test technique measure? What the C-test is measuring is still open to question. There have always been two sides on this issue. On the one hand are those who argue that C-tests assess more grammatical competence than textual competence. As seen in the discussions above, the C-test only requires the student to fill in second halves of words. In completing a given word, the most important clues for the test taker are in the immediate environment of the blank (Klein-Braley, 1985), including the first half of the word itself. On the basis of student performance, researchers have found that recognition of syntactical relationships comes first, after which understanding of semantic relationships is necessary for the true comprehension necessary for good performance. Simply stated, the students probably look first for clues in the half of the word that is given; then after formulating a guess, they gradually look for the relationship in meaning between their guess and the words around it.

Experiments with C-tests have primarily involved correlating C–test scores with other types of language tests. For example, the Chapelle and Abraham (1990) study revealed that C-tests correlated strongly with a vocabulary test. On the basis of this, the conclusion they reached was that C-tests tend to measure the grammatical competence of students. On the other hand, validity research suggests that the C-test is a measure of overall language proficiency (Stansfield & Hansen, 1983). The C-test was explicitly developed as a test of general language proficiency (Klein-Braley and Raatz, 1984). Indeed, some results obtained with C-tests show meaningful relationships with other tests of general language knowledge and performance. Still other experiments reveal that C-test scores increase regularly and predictably with an individual's linguistic maturational level (Klein-Braley, 1985). This notion is further supported by the fact that adult native speakers perform well on C-tests.

When and How to Use the C-test

C-tests have been shown to be both empirically and theoretically valid. It is now time to go back to the more practical issues involved in classroom testing. Being an objective test type, the most important characteristics of a C–test are the fact that it is easy to construct and easy to score. Depending on the length of the test, it can take 30-45 minutes to administer, at least at the high school level.

C–tests are, however, typically very difficult: "one expects that on average only half of the mutilations will be correctly restored"

(Klein-Braley & Raatz, 1984). A test with a 50% average may be very frustrating for both the students and teacher. Thus the C-test should not be used at the end of a course unit where mastery of the subject is required and when an average of about 80% may be desired. The main usefulness of the C-test lies at the start or end of a course, when it is necessary to determine the ranking of students in relation to each other within a group. In addition, the empirical research indicates that C–tests are grammatically based. Hence, teachers might want to make use of this test to measure the students' grammatical competence and perhaps vocabulary development. Both teachers and language administrators can use the classroom C-test results with other more complex language tests in selecting and placing students in a language program. The section above on what skills the C-test measures cited some evidence that C-test performance increases with the student linguistic maturational level. Thus the language teacher can use the results of a C-test to measure and compare periodic language skills progress of the same group of students. This may sound a little too big a task, but given the relative ease with which the C-test can be constructed, administered, and scored, it may prove worth trying.

Conclusion

This chapter has outlined four different types of cloze in relation to classroom testing. Cloze procedures hold potential for measuring several different aspects of students' second language competence (Bachman, 1990). The specific language traits measured by a particular cloze depends on the methods of cloze construction and on the types of responses required of students.

The first type, the fixed-ratio cloze, is intended to sample various types of words on a regular basis; some words may be governed by local, grammatical constraints, and others, by long-range, textual constraints. The second type, the rational–deletion cloze, allows the test developer control over the types of words deleted, and thus the language traits measured. The third type, by altering the mode of expected response, requires the students to select the correct answer from a given set of choices. In the fourth type, the C-test, deletions are made on the second half of every other word. Because of the shorter segments of text and the importance of clues in the immediate environment (Chapelle & Abraham, 1990), this procedure most likely results in a test that is more directly related to grammatical competence.

While debate within the empirical research on cloze testing still continues, the best that second language teachers can do is to apply whatever research can offer so far.

References

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a second language. *TESOL Quarterly, 13*, 219-226.

Alderson, J. C. (1983). The cloze procedure and proficiency in English as a second language. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 205-217). Rowley, MA: Newbury House.

Anderson, J. (1976). *Psycholinguistic experiments in foreign language testing*. St. Lucia, Queensland, Australia: University of Queensland Press.

Bachman, L. (1982). The trait structure of cloze test scores. *TESOL Quarterly, 16*, 61-70.

Bachman, L. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly, 19*, 535-556.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University.

Bensousan, M., & Ramraz, R. (1984). Testing reading comprehension using a multiple-choice rational cloze. *Modern Language Journal, 68*, 230-239.

Brown, J. D. (1983). A closer look at cloze: Validity and reliability. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 237-250). Rowley, MA: Newbury House.

Brown, J. D. (1985). Cloze procedure: A tool for teaching second language reading. *TESOL Newsletter, 20*(5), 1 & 7.

Brown, J. D. (1988a). Tailored cloze: Improved with classical item analysis techniques. *Language Testing, 5*(1), 19-31.

Brown, J. D. (1988b). What makes a cloze item difficult? *University of Hawaii working papers in ESL, 7*(2), 17-39.

Brown, J. D. (1991). What test characteristics predict human performance on cloze test items? In the *Proceedings of the Third Conference on Language Research in Japan* (pp. 1–26). Urasa, Japan: International University of Japan.

Chapelle, C. (1988). Field independence: A source of language test variance? *Language Testing, 5*, 62-82.

Chapelle, A., & Abraham, R. (1990). Cloze method: What difference does it make? *Language Testing, 7*, 121-146.

Chavez-Oller, M. A., Chihara, T., Weaver, K. A., & Oller, J. W., Jr. (1985). When are cloze items sensitive to constraints across sentences? *Language Learning, 5*, 62-82.

Hale, G., Stansfield, C., Rock, D., Hicks, M., Butler, F., & Oller, J. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing, 6*, 49-78.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hanania, E., & Shikhani, M. (1986). Interrelationships among three tests of language proficiency: Standardized ESL, cloze and writing. *TESOL Quarterly, 20*, 97-109.

Hinofotis, F., & Snow, B. (1980). An alternative cloze testing procedure: multiple-choice format. In J. W. Oller, Jr., & K. Perkins (Eds.), *Research in language testing* (pp. 238-245). Rowley, MA: Newbury House.

Ikeguchi, C. (Unpublished ms.) The four cloze test types: To each its own. Dokkyo University, Soka–shi, Saitama Prefecture, Japan.

Jonz, J. (1976). Improving on the basic egg: The M–C cloze. *Language Learning, 26*, 255-265.

Jonz, J. (1990). Another turn in the conversation: What does cloze measure? *TESOL Quarterly, 24*, 61–83.

Klein-Braley, C. (1983). A cloze is a question. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 134-156). Rowley, MA: Newbury House.

Klein-Braley, C. (1985). A cloze-up on the C test. *Language Testing, 2*, 76-104.

Klein-Braley, C., & Raatz, U. (1984). A survey of research on the C-test. *Language Testing, 1*, 134-146.

Lo Castro, V. (1994). Teachers helping themselves. *The Language Teacher, 18* (2), 4-7.

Mochizuki, A. (1994). Four kinds of texts, their reliability and validity. *JALT Journal, 16*, 41-54.

Ozete, P. (1977). The cloze procedure: A modification. *Foreign Language Annals, 10*, 565-568.

Perkins, J., & German, P. (1985). The effect of structure gain on different structural category deletions in a cloze test. Paper presented at Midwest TESOL. Milwaukee, WI.

Porter, D. (1983). The effect of quantity and quality on the ability to make predictions. In D. Porter, & A. Hughes (Eds.), *Current developments in language testing*. London: Academic Press.

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17*, 229-255.

Stansfield, C., & Hansen, J. (1983). Field dependence-independence as a variable in second language test performance. *TESOL Quarterly, 17*, 29-38.

Taylor, W. L. (1953). Cloze procedure a new tool for measuring readability. *Journalism Quarterly, 30*, 415-433.

Yamashita, S. (1994). Is the reading comprehension performance of learners of Japanese as a second language the same as that of Japanese children? An analysis using a cloze test. *Sekai no Nihongo Kyōiku, 4*, 133-146.

Chapter 18

The Validity of Written Pronunciation Questions: Focus on Phoneme Discrimination

SHIN'ICHI INOI
OHU UNIVERSITY

Pronunciation questions have been so popular that they are often included in entrance examinations to colleges and universities. They are even found in Center Exams administered by the Ministry of Education. Some researchers, however, have recently cast doubt on the validity of pronunciation questions on written tests as a means of evaluating students' actual pronunciation ability. Among these critics are Katayama, Endo, Kakita, and Sasaki (1985), Takei (1989a, 1989b), and Wakabayashi and Negishi (1993). These critics all claim that the teacher cannot assess students' pronunciation performance with questions on a written test and that such questions cannot be a substitute for an oral test. They all stress the importance of an oral test in evaluating students' pronunciation ability. However, Sasaki and Tomohiko (1991) and Shirahata (1991), took a somewhat different view on this point. While they admitted the low-content validity of phoneme discrimination questions on their written test, they claimed the validity of primary stress questions.

In a previous study (Inoi, 1994), I investigated whether pronunciation questions on written tests are a valid measurement of students' pronunciation ability. A written test of 20 questions on primary stress and another 20 questions on phoneme discrimination were administered to 44 Japanese college freshmen in a language laboratory. After the written test, an oral version was given to the same subjects. The written primary stress questions were all of the same type: the subjects were to choose which syllable of the word in question had primary stress. The written phoneme discrimination test, divided into two sections, had the same format as in the present study (see Appendix A). On the oral version of the primary stress test, the subjects were instructed to pronounce each of the words. On the oral phoneme discrimination test, the subjects were asked to pronounce all the words on the items, including alterative responses. The subjects' pronunciations on both primary stress and phoneme discrimination tests were recorded on cassette tapes. The scoring criteria employed for the oral data, which were the same as those for the present study, are explained in detail later. The data obtained from the written test and the oral test were analyzed in terms of the agreement rate, that is, the extent to which the answers were identical between the two tests, as well as in terms of the correlation between the scores on the written and oral tests.

Figure 1. *Contingency Table*

		Oral Test	
		Correct	Wrong
Written Test	Correct	A	B
	Wrong	C	D

To calculate a subject's agreement rate, each subject's answers on the two tests were sorted into a two-by-two contingency table, as shown in Figure 1. Cell A shows the number of answers correct on both the written and oral tests. Cell B shows the pairs of answers correct on the written test but wrong on the oral test. Cell C shows the pairs correct on the oral test but wrong on the written test, and D shows the pairs wrong on both tests. The agreement rate was calculated as follows: the sum of A and D were divided by the total (i.e., A + B + C + D), then the value obtained was multiplied by 100.

For the primary stress test, the average agreement rate of the subjects was 77.9 percent, or about 80 percent. A moderate correlation was observed between scores on the two tests ($r = .65$, $df = 42$, $p < .01$). Therefore, it was concluded that the primary stress questions on the written test were reasonably sound measures.

As for phoneme discrimination questions, the average agreement rate was 67.7 percent, about 10 percent lower than the one for primary stress questions. In one of the two sections of the test, a moderate correlation was found ($r = .61$, $df = 42$, $p < .01$), but in the other section, no statistically significant correlation was observed ($r = .22$, $df = 42$, NS). Thus, compared with the primary stress test, the phoneme discrimination test was relatively weak. The non-significant correlation obtained indicated that the phoneme discrimination in that section may not be valid.

However, each of the two sections of the phoneme discrimination test contained only 10 questions and the questions used were randomly taken from various college entrance examinations. There remains the possibility that a greater number of questions and a greater variety of words used for the questions may produce results different from those in my previous study. The present study is an attempt to pursue that possibility. By focusing on phoneme discrimination questions, this study further investigates the effectiveness of a written phoneme discrimination test.

Experimental Procedure

The experiment was administered to 60 college freshmen in the language laboratory of Ohu University in Koriyama, Fukushima-ken,on July 18, 1994. The subjects were all Japanese learners of English as a foreign language. As in the previous study, a written test was administered first, then it was followed by an oral version. Each subject was given a test sheet of 30 phoneme discrimination items. The test was multiple-choice and was divided into two parts (Section A and Section B), each of which had 15 items. In Section A, the first five items were taken from the 1994 Center Exam, the next five from the 1993 exam, and the last five from the 1992 exam. In Section B, the first five were from the 1989 exam, the second five from the 1989 supplementary exam, the next four from the 1988 exam, and the last one from the 1988 supplementary exam (see Appendix). No use of dictionaries was allowed. All of the subjects finished the test in less than 20 minutes. After the written test, each subject was given another answer sheet. It was the same as the one used for the written test. The subjects were instructed to read aloud each of the words on the items. Their pronunciations of the words were recorded on tape. Both the written tests and the oral recordings were scored by the author.

The experimental procedure and the data-scoring criteria were the same as those employed in the earlier study. I wanted to compare statistical results with those found in my previous study. The data were analyzed in terms of the correlational relationship of the written and oral test scores and in terms of the agreement rate of answers between the two.

The agreement rate of answers was calculated for each subject and each item. The subjects' agreement rates were examined further to see whether they had any correlation with subjects' accuracy scores. Based on the results, the validity of phoneme discrimination questions on a written test will be discussed along with some of the problems they pose.

Data Analysis

Each subject produced one pair of answers for each test item (i.e., one answer on the written test and another on the oral test). In total, each subject produced 30 pairs of answers: 15 pairs in Section A and another 15 in Section B. As in the previous study, the scoring of the oral test was based on the pronunciation of the whole word. This was done because, in general, teachers seem to pay most attention to the pronunciation of a word as a whole rather than to its parts when they evaluate how a student pronounces English words. The written test was only concerned with the underlined parts of words as indicated in the instructions. In Section A, each subject was given a score when the subject correctly pronounced both the head word (i.e., the first word on the left in each item) and the correct option.[1] However, when either of the two words was pronounced with primary stress on the wrong syllable, a score was not given. On item 15, for instance, when the head word "dessert" was pronounced as [dézərt], the answer was not judged to be correct. In section B, each subject was given a correct score when the subject pronounced all four options correctly. When any one of the four options was pronounced with stress on the wrong syllable, the answer was counted as incorrect.

Results and Discussion

The reliability of the written and oral tests was estimated by using Kuder-Richardson Formula 20 (Brown, 1988). Table 1 shows the reliability coefficients for each section of the written and oral tests as well as for the test as a whole. Relatively high reliabilities of .79 and .86 were obtained for the whole written test and for the whole oral test, respectively.

Table 1. *Reliability estimates of the written and oral tests*

Test	Section	K-R20
Written	A	.64
	B	.69
	A+B	.79
Oral	A	.68
	B	.75
	A+B	.86

Correlational Analyses

Tables 2, 3, and 4 indicate both the individual and combined results of the correlational analyses of Sections A and B. In each case, a fairly strong correlation was observed between written test scores and oral test scores. As shown in the tables, the correlation coefficient for Section A was .73, that for Section B was .74, and that for both Sections A and B was .81. These figures are considerably higher than those observed in the previous study, where the correlation coefficient of the section corresponding to Section A was .61, that for the section corresponding to Section B was only .22, and the overall coefficient was .51. Thus, in the previous study, I doubted the validity of phoneme discrimination items in the section where no significant correlation was observed. In the present study, however, the stronger correlations that were observed generally support the validity of phoneme discrimination on the written test. But analyses of agreement rates below show that there may still be some problems with the validity of the written test.

Agreement Rate Analyses

Agreement rates of answers on the written and oral tests were analyzed from two different points: for each subject and each item.

Subjects' agreement rates. In calculating a subject's agreement rate, the same procedure was employed as in the previous study. Table 5 indicates the agreement rates of each

Table 2. *Pearson Product-Moment Correlation of scores between the written and oral tests on Section A*

	N	Mean(%)	S	r
Written Test	60	47.8	19.0	.73*
Oral test	60	46.6	19.5	

*$p < .01$

Table 3. *Pearson Product-Moment Correlation of scores between the written and oral tests on Section B*

	N	Mean(%)	S	r
Written Test	60	68.2	18.4	.74*
Oral test	60	46.3	20.6	

*$p < .01$

Table 4. *Pearson Product-Moment Correlation of scores between the written and oral tests on Sections A and B*

	N	Mean(%)	S	r
Written Test	60	48.0	16.9	.81*
Oral test	60	47.4	19.0	

*$p < .01$

subject's answers between the written and oral tests. The average agreement rates for Section A, Section B,and both Sections A and B combined were 65.3, 68.3, and 66.8 percent, respectively. The overall agreement rate of 66.8 percent was almost the same as the 67.7 percent found in the previous study.

An analysis was done to detect any correlation between the subjects' agreement rates and their accuracy scores. As shown in Table 6, a weak but statistically significant correlation of .37 was observed between agreement rates and written test scores; a moderate and statistically significant correlation of .56 was found between agreement rates and oral test scores.

These correlations can be more clearly seen in Figure 2. It shows the accuracy scores on the written and oral tests attained by those subjects who belonged to four different agreement rate categories. The abscissa shows the agreement rate categories and the ordinate represents accuracy scores. The figures in parentheses represent the number of subjects who fall into each of the four agreement rate categories. The first pair of bars in Figure 2 indicates that, for those subjects whose agreement rate averaged 80 percent or more, their average accuracy score was 81.9 percent on the written test and 70.5 percent on the oral test. There is a clear tendency for high agreement rate achievers to have high accuracy scores and for low agreement rate achievers to have low accuracy scores on the tests. However, this tendency is less clear when comparing those subjects in the lowest rate category with those in the second lowest category.

This relationship between agreement rates and accuracy scores might be accounted for by the following: high accuracy score achievers may have known many of the English words used on the tests, which enabled them to choose the correct option on the written test and to pronounce the words correctly on the oral test. Low accuracy score achievers, on the other hand, may have had less knowledge of the test words. Thus, on the written test, low accuracy score achievers may have simply chosen answers through random guessing, and, consequently, on some items may have select the correct answer. On the oral test, however, these subjects may often have been unable to pronounce the words correctly, and they were less able to provide correct answers by chance alone. In this way, high agreement rates would not be found between the answers given on both tests by the subjects with low accuracy scores. In other words, phoneme discrimination items used on the written test may not accurately assess low accuracy score achievers' actual pronunciation performance.

Agreement rates of items. Table 7 shows the agreement rates of items, or the extent to which the subjects' answers on each item were the same between the written and oral tests. On item 1, for example, the answers given by 46 subjects, or 76.7 percent of the total, were identical between the written and

Table 5. *Agreement rates of each subject's answers between the written and oral tests*

	SECTION				SECTION		
Subject	A	B	A+B	Subject	A	B	A+B
S1	66.7	80.0	73.3	S31	66.7	73.3	70.0
S2	66.7	73.3	70.0	S32	46.7	60.0	53.3
S3	53.3	73.3	63.3	S33	40.0	66.7	53.3
S4	93.3	80.0	86.7	S34	73.3	66.7	70.0
S5	66.7	60.0	63.3	S35	60.0	53.3	56.7
S6	60.0	80.0	70.0	S36	53.3	73.3	63.3
S7	73.3	73.3	73.3	S37	60.0	53.3	56.7
S8	73.3	66.7	70.0	S38	46.7	73.3	60.0
S9	80.0	80.0	80.0	S39	60.0	66.7	63.3
S10	40.0	60.0	50.0	S40	46.7	46.7	46.7
S11	66.7	60.0	63.3	S41	66.7	86.7	76.7
S12	66.7	73.3	70.0	S42	66.7	80.0	73.3
S13	73.3	73.3	73.3	S43	73.3	60.0	66.7
S14	73.3	73.3	73.3	S44	66.7	73.3	70.0
S15	73.3	86.7	80.0	S45	66.7	73.3	70.0
S16	73.3	73.3	73.3	S46	53.3	80.0	66.7
S17	73.3	80.0	76.7	S47	66.7	53.3	60.0
S18	66.7	80.0	73.3	S48	73.3	60.0	66.7
S19	60.0	73.3	66.7	S49	73.3	60.0	66.7
S20	40.0	66.7	53.3	S50	93.3	73.3	83.3
S21	53.3	60.0	56.7	S51	60.0	73.3	66.7
S22	73.3	93.3	83.3	S52	73.3	46.7	60.0
S23	60.0	73.3	66.7	S53	73.3	66.7	70.0
S24	80.0	80.0	80.0	S54	60.0	53.3	56.7
S25	73.3	66.7	70.0	S55	46.7	73.3	60.0
S26	73.3	46.7	60.0	S56	66.7	60.0	63.3
S27	53.3	53.3	53.3	S57	80.0	73.3	76.7
S28	80.0	73.3	76.7	S58	66.7	93.3	80.0
S29	66.7	53.3	60.0	S59	66.7	73.3	70.0
S30	53.3	53.3	53.3	S60	66.7	33.3	50.5
				AVERAGE	65.3	68.3	66.8

the oral tests. On three items, items 15, 19, and 20, the agreement rates failed to reach 50 percent.

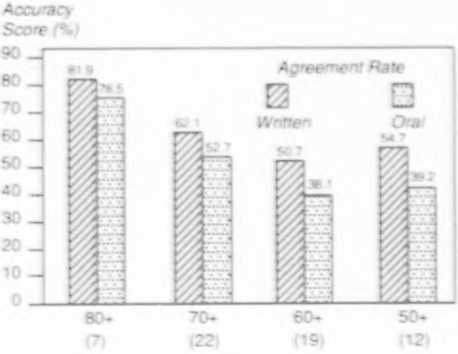
A number of factors may have affected the low agreement rates on each of these items. On item 15, there were many subjects who chose the correct option on the written test but were unable to correctly pronounce either the head word "dessert" or the correct option "possess" on the oral test: the head word was incorrectly pronounced as [dézərt], with primary stress on the wrong syllable; the correct option was pronounced as [p zes], [póuziz], or [pəzíz]. As revealed in their oral recordings of these words, these subjects pronounced the phonemes as [z], and yet, on the written test, because they judged that both the phoneme in question (i.e., the underlined part of the head word) and the one in the correct option had the same pronunciation, they were able to select the appropriate answer. In fact, there were 25 such subjects, or 41.7 percent of the total, which could have led to the low agreement rate on this item.

Table 6. *Pearson-Product Moment Correlation between agreement rates and accuracy scores*

	N	Mean(%)	S	r
Agreement rate	60	66.8	9.2	.81*
Accuracy score on the written test	60	58.0	16.9	.37*
Accuracy score on the oral test	60	47.4	19.0	.56*

*$p < .01$

Figure 2. *Correlational Relationship between Areement Rate and Accuracy*



On item 19, the low agreement rate was also caused by some subjects' inconsistent performance between the written and oral tests. As many as 25 subjects were not able to correctly pronounce the "bathe" option on the oral test, while they all chose this correct option on the written test. A close analysis of their oral recordings of the words on the item shows that they were able to pronounce correctly all the options except the "bathe." It seems that the distractors in this item were a too easy for these subjects so they could eliminate the options as possible answers even if they did not know how to pronounce the correct answer. In short, the inclusion of easy words as distractors may have caused the low agreement rate on item 19.

On item 20, as many as 22 subjects, about one-third of the total, were able to pronounce all four options correctly, but somehow they chose a wrong option as the answer on the written test. At the moment, no reasonable explanation can be given for these results.

Since the above three items did not seem to accurately assess subjects' oral performance, their validity is in doubt.

Conclusion

The present study, a follow-up to my earlier study, addressed the issue of the validity of pronunciation questions on a written test for assessing the pronunciation of English words. In the previous study, phoneme discrimina-

Table 7. *Agreement rates of answers on items between the written and oral tests*

Item	1	2	3	4	5	6	7	8	9	10
N*	46	38	38	38	34	41	42	35	43	38
AR(%)**	76.7	63.3	63.3	63.3	56.7	68.3	70.0	58.3	71.7	63.3
Item	11	12	13	14	15	16	17	18	19	20
N	46	35	44	44	24	37	32	41	29	27
AR(%)	76.7	58.3	58.3	73.3	40.0	61.7	53.3	68.3	48.3	45
Item	21	22	23	24	25	26	27	28	29	30
N	32	43	37	48	49	52	54	41	48	45
AR(%)	53.3	71.7	61.7	80.0	81.7	86.7	90.0	68.3	80.0	75.0

*N = Number of subjects whose answers were identical between the written and oral test

**AR = Agreement rate

tion questions were taken from various college entrance examinations, and no significant correlation was observed between scores on the written and oral sections of the test. As a result, for this study, I chose to use phoneme discrimination questions taken from *Center* examinations administered by the Ministry of Education.

Unlike the earlier study, the results of the present study showed a relatively strong correlation between scores on the written and oral tests in each of the two sections (Section A and Section B). From a correlational point of view, the statistically significant correlations supported the validity of phoneme discrimination items on the written test.

From an agreement-rate point of view, however, their validity was not so strongly supported. Though the subjects' average agreement rate between answers on the written and oral tests reached nearly 70 percent for each of the two sections, an analysis of the agreement rates showed that there was a tendency for high agreement rate achievers to get high accuracy scores and for low agreement rate achievers to get low accuracy scores. As for subjects with low accuracy scores, or presumably low-level students, phoneme discrimination items on the written test did not appear to very accurately assess their actual performance on the oral test and some doubt should be cast on the validity of such questions.

An analysis of the agreement rates on the items showed that there were three items whose agreement rates were below 50 percent. An examination of the subjects' answers on these items revealed some problems in the written test. One problem was that even if subjects did not know how to pronounce correctly either the head word or the correct option on one item in Section A, there was some chance for them to choose the correct answer on the written test. Another problem was that the distractors on one item in section B were so easy that subjects could eliminate them as possible answers. Still another problem was the fact that the good performance by some subjects on the oral test did not necessarily mean they would perform well on the written test. These problems indicate that some phoneme discrimination items on the written test were not valid means of assessing the subjects' actual pronunciation ability of English words.

In making phoneme discrimination questions on a written test with a multiple-choice format, teachers should bear in mind the following points:

1. The distractors on an item should be carefully designed lest they be eliminated too easily.
2. Teachers may not be able to very accurately assess the pronunciation ability of low-level students with a written test.
3. Good performance on a written test does not always guarantee good performance on an oral test.

Suggestions for Future Research

The results of the present study were based on data collected from a written test of 30 phoneme discrimination questions and their oral versions. While the reliability estimate was, as a whole, relatively high for both the written and oral tests, it was not so high for each section of the test, particularly for Section A of the written test. This was probably because each section contained only 15 items. A test with a greater number of items is necessary to improve the reliability of a written phoneme discrimination test. Results may change for the same items when different samples of subjects are used, or when different test formats or scoring criteria are used. Results may also change depending upon learners' English proficiency levels. Future research should examine these points to further investigate the validity of pronunciation questions on a written test.

Note

[1] Of course, there may be other ways of scoring the oral data. But I wanted to employ the same scoring procedure as the one used in the previous study.

References

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design.* New York: Cambridge University Press.

Henning, G. (1987). *A guide to language testing: development, evaluation, and research.* Rowley, MA: Newbury House.

Inoi, S. (1994). A study of the validity of pronunciation questions on a written test. *JACET Bulletin, 25,* 39-52.

Katayama, Y., Endo, H., Kakita, N. & Sasaki, A. (1985). *Shin-eigokakyoiku no kenkyu.* Tokyo: Taishukan.

Kyogakusha-shuppan senta. (1994). *Daigakunyushisentashiken mondaikenkyu.* Tokyo: Kyogakusha.

Sasaki, C. & Tomohiko, S. (1991). An investigation into the validity of paper test problems on stress and phonemes. *Annual Review of English Language Education in Japan, 3,* 119-127.

Shirahata, T. (1991). Validity of paper test problems on stress: Taking examples from Mombusho's daigaku nyushi senta shiken. *Bulletin of the Faculty of Education, Shizuoka University, Educational Research Series, 23,* 161-172.

Takei, A. (1989a). Paper and pencil tests for pronunciation: Are they valid? *Bulletin of the Kanto-Koshinetsu English Language Education Society, 2,* 17-22.

Takei, A. (1989b). More on the validity of paper and pencil tests for pronunciation. *The IRLT Bulletin, 3,* 1-21.

Wakabayashi, S. & Negishi, M. (1993). *Musekininna tesutoga ochikobore wo tsukuru.* Tokyo: Taishukan.

Appendix

A. Choose the correct option whose underlined part is pronounced the same as that of the first word on the left.

1. wilderness a. hive b. myth c. theme d. triumph
2. raw a. coast b. naughty c. notice d. route
3. chimney a. action b. cheminstry c. natural d. scholar
4. passion a. assure b. blossom c. confess d. scissors
5. conquer a. conquest b. liquid c. quiet d. unique
6. mood a. flood b. floor c. shoot d. wool
7. dear a. beard b. heart c. pearl d. wear
8. country a. although b. doubt c. southern d. though
9. recent a. ancient b. decorate c. financial d. society
10. language a. argument b. distinguish c. guess d. regular
11. allow a. bowl b. coward c. grow d.knowledge
12. imagine a. capital b. false c. major d. sacred
13. brush a. bury b. bush c. rude d. thumb
14. rough a. brought b. cough c. ghost d. thorough
15. dessert a. assume b. message c. permission d. possess

B. Choose the correct option whose underlined part is pronounced differently from the three other examples.

16. a. control b. hostess c. improve d. postcard
17. a. breath b. creature c. feather d. treatment
18. a. horizon b. isolate c. polite d. risen
19. a. bathe b. both c. thirsty d. thousand
20. a. accident b. account c. accuse d. accustom
21. a. behavior b. canal c. label d. parade
22. a. heaven b. pleasant c. steady d. stream
23. a. family b. reply c. ugly d. weekly
24. a. arch b. March c. speech d. stomach
25. a. loose b. news c. poison d. resemble
26. a. coffee b. joke c. nose d. smoke
27. a. delight b. describe c. mild d. wisdom
28. a. bear b. dear c. fear d. near
29. a. chemistry b. chimney c. chorus d. mechanical
30. a. eight b. height c. neighbor d. weight

14 Reasons Why You Should Join The Japan Association of Language Teachers

Compiled by Don Modesto

1 Leading authorities in language teaching regularly visit us: H. Douglas Brown, John Fanselow, Jack Richards, Kathleen Graves, Alan Maley... (If you don't know who they are, come to JALT to find out.)

2 Insights on the job market, introductions... JALT plugs you into a network of over 3,600 language teaching professionals across Japan.

3 Ten special interest groups and their newsletters (and more of each on the way): Bilingualism, College and University Educators, Computer Assisted Langauge Learning, Global Issues in Language Education, Japanese as a Second Language, Learner Development, Materials Writers, Teacher Education, Team Teaching, and Video.

4 JALT is a place to call your professional home. And with 38 chapters across Japan, JALT is not far from your other home.

5 Monthly chapter programs and regular regional conferences provide both valuable workshops and the chance to share ideas and hone your presentation skills.

6 Professional organizations look great on a resumé. Volunteer for a chapter position, work on a conference, or edit for the publications. You gain management and organizational skills in the bargain.

7 JALT maintains links with other important language teaching organizations such as TESOL and IATEFL. We have also forged links with our counterparts in Korea and Taiwan.

8 Publishing your research? Submit it to the JALT *Journal*.

9 Looking for a regular source of language teaching tips? Look to our celebrated magazine *The Language Teacher*—the only language teaching publication in the world published on a monthly basis.

10 JALT produces Asia's largest language teaching conference, with dozens of publishers displaying the latest materials, hundreds of presentations by leading educators, and thousands of attenders.

11 JALT nurtures a strong contingent of domestic speakers: Dale Griffee, Marc Helgesen, Kenji Kitao and Kathleen Kitao, Don Maybin, Nishiyama Sen, and Wada Minoru.

12 Conducting a research project? Apply for one of JALT's research grants.

13 Free admission to monthly chapter meetings, reduced conference fees, subscriptions to *The Language Teacher* and *JALT Journal* and this for just ¥7,000 per year for individual membership, ¥6,000 for joint (2 people), and ¥4,500 if you hustle a little and get up a group of four to join with you.

14 Easy access to more information, application procedures, and the contact number of the chapter nearest you.

Contact JALT's Central Office today.
03-3802-7121 (phone) or
03-3802-7122 (fax)

ISBN: 4-9900370-0-6 ¥2500

Language Testing in Japan is the first collection of articles in the JALT Applied Materials series, as well as the first contemporary collection of articles in English devoted to the issues of testing, assessment and evaluation in Japan. Written by classroom teachers in Japan, *Language Testing in Japan* offers chapters on classroom testing strategies, program-level testing strategies, standardized testing, oral proficiency testing, and innovative testing, while at the same time providing information about actual testing that is currently being done in many of the language programs and classrooms throughout Japan. All of the chapters have been written to make them as practical, useful, straightforward, and easy to understand as possible.

Standardized tests such as TOEFL, TOEIC, SPEAK, and STEP are discussed and their strengths and weaknesses analyzed. Other issues addressed are university entrance exams, cloze tests and how to make them, how to make and improve teacher-made classroom tests, how to test young learners, whether paper pronunciation tests are valid, and whether it is possible to test nonverbal abilities. Technical testing terms such as *norm-referenced* and *criterion-referenced* tests, *item facility, skewdness, normal curve, washback*, and so on are explained in detail.

Edited by the internationally recognized language testing expert James Dean Brown and Japan testing expert Sayoko Okada Yamashita, and written *by* language teachers *for* language teachers, *Language Testing in Japan* assumes no prior knowledge of language testing yet makes a long-needed contribution to this much discussed area. It is certain to stimulate discussion for years to come about language testing as it is practiced today in Japan.

A Special Supplement to The Language Teacher

BEST COPY AVAILABLE